JIGSAW 0.53 Tutorial
January 2014

Table of Contents

0. Overview and Purpose 2
1. Getting Started with Jigsaw 3
2. Importing and Saving Documents 7
2.1 Importing Documents 7
2.2 Jigsaw Projects and Workspaces 10
3. Identifying and Working with Entities 11
3.1 Entity Identification 11
3.2 Correcting Erroneous Entity Identification 13
3.3 Manipulating Entities 13
3.4 Entity Aliasing 14
4. Exploring and Analyzing a Document Collection 15
4.1 General Tips 15
4.2 Search Tips 16
4.3 View-specific Use Tips 17
4.4 Automated Computational Analysis 22
4.5 Gathering Evidence with the Tablet 24
5. Additions to Distribution 26
6. Known Issues/Bugs 28
7. Help/Comments 29
8. Coming Soon 30
Appendix 31

0. Overview and Purpose

Jigsaw is a visual analytics system designed to help people browse, explore, analyze, understand and make sense of collections of text documents. Jigsaw presents multiple visualizations of the documents and the entities within them, with a special focus on showing connections between entities (entities that appear together in some document).

Jigsaw is designed to work best with collections of many documents that are relatively short in length. By many documents, we mean collections that may go up to 5,000-10,000 documents. We think that Jigsaw possibly could run well on collections approaching 50,000 or 100,000 documents too, given that your computer has very fast processors and a great deal of RAM. The documents ideally should be about 1-6 paragraphs in length, that is, about a page or two. What is most important here is the number of named entities per document. This number should likely be below about 50-75 entities for Jigsaw to be most helpful.

Jigsaw is *not* designed to analyze a small number of extremely large documents like books or academic papers. These types of documents should be broken into smaller units such as sections, subsections, or pages, and then each of these units becomes its own document.

Because Jigsaw provides many different visualizations of the documents and entities, you should ideally have a large amount of screen real estate to show the views. We find that the system functions best on a computer with a large display such as a 30" monitor or on a computer with multiple monitors. While you can still run the system on a single smaller monitor, you may be limited in the number of views you can easily manipulate or you may have to do much window swapping via the taskbar or other means.

We have tried to keep this Tutorial relatively concise so that you can easily read and browse it, while still communicating the most important information necessary to effectively use the system.

This documentation is also available on-line in HTML format at http://www.cc.gatech.edu/gvu/ii/jigsaw/tutorial/manual. If you prefer interactive video assistance over a printed document (and who doesn't these days), you should watch the tutorial videos on the web page http://www.cc.gatech.edu/gvu/ii/jigsaw/tutorial. They should be helpful in learning how to use the system. The "System Requirements" video, in particular, provides assistance with getting the system set up and running on your machine.

1. Getting Started with Jigsaw

This section will help you to quickly familiarize yourself with how to run Jigsaw.

1.0 System requirements

We have tested Jigsaw on Windows XP, Windows Vista, Windows 7, and 8 platforms, on Linux, and on a Macintosh running MacOS 10.5-10.9. It should run well on all of these platforms. You can run Jigsaw on a machine with 1 GB of RAM but we don't recommend it. You should really have at least 2 GB of RAM to run the system. Having more than that is even better.

You must have the Java Version 6 (or higher) installed in order to run Jigsaw. To check what version of Java is being used on your machine, go to a command shell/terminal/window and type the command "java –version" (do not include the quotation marks). If the second digit in the reported Java version is a '6' or higher, then you are OK. If it is a '5' or a '4', you need to update to a newer version of Java.

If you are on a Windows machine and need to update to a newer version of Java, it can be downloaded at http://www.java.com. This is the same for Macs with OS 10.7 or newer.

On a Mac with OS 10.6, you do not load a newer version of Java by going to the website above. Instead, you should first do a System Update to download version 6 of Java. You next must run the "Java Preferences" command in the Utilities folder of the Applications folder. There, change the order of the bottom region (Java applications) so that Java version 6 is the top item in the list.

For advanced users: We have configured the Jigsaw start-up command to request 1 GB of memory for the Java virtual machine. If you are running on a machine with more RAM, you can bump up this number. To do so, simply edit either the Jigsaw.bat (Windows), Jigsaw.command (Mac), or Jigsaw.sh (Linux/UNIX) script file and change the two occurrences of 1024 (which provides 1 GB) to something like 2048 or 4096.

1.1 Initiating a session

On a Windows machine, double-click on the `Jigsaw.bat` script file in the Jigsaw folder to start the system. This is a short script file that will start Java and will try to give the system a good amount of heap memory. If this script runs successfully, then it will bring up the Jigsaw Control Panel, shown below.

On a Mac, double-click on the `Jigsaw.command` script file in order to start the system. Note that under Mac OS 10.8 (Mountain Lion) or newer, you cannot run it by simply double-clicking on the Jigsaw.command file because it was downloaded from the internet. In this case, you need to right mouse button-click (or control-click) in order to bring up a menu and then you choose the "Open" option. Once you do that, you will be able to just use a double-click on subsequent tries.

On Linux/UNIX, execute the shell command file `Jigsaw.sh`.

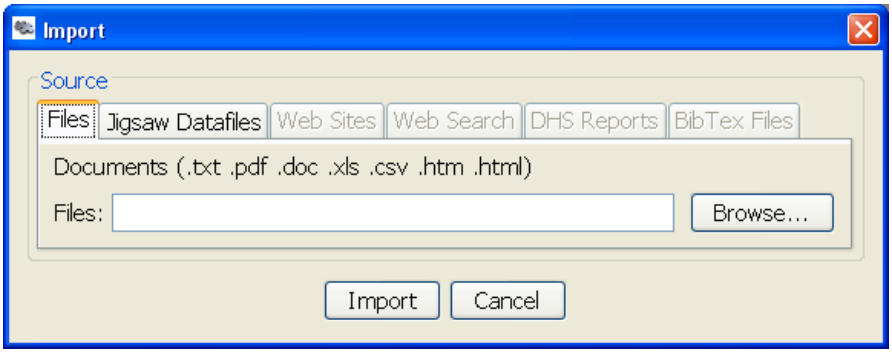
1.2 **Reading in a Set of Documents**

Jigsaw can read in (and store) documents from a variety of formats. It can read original documents such as text, csv, html, pdf, Word, and Excel files. We also have created a Jigsaw Datafile format using xml that can be read in. Additionally, there are a few specific, proprietary document formats that Jigsaw can read.

To import a source document that has not been processed at all yet, use the *File* menu's *Import* command. This will bring up the *Import* dialog box with tabs for the different types of documents that can be read in.

The main tab here is *Files*. It allows you to read in plain text (.txt), MS Word (.doc), PDF (.pdf), html (.htm or .html), comma-separated value (.csv), and MS Excel (.xls) files. To read in multiple files at once, use the *Browse* button and select multiple files in the chooser dialog box. For csv and MS Excel files, a special mapping process will begin that allows you to specify what each column in the file means (more on that later in this document). We hope to soon add the ability to read in pages and sites from web crawls, pages from web searches, and bibliographic style pages.

The files you import can be simple ascii text or they can be Unicode. Because Jigsaw can now read Unicode, text from international (non-English) languages can be handled in Jigsaw too.



A second tab in the *Import* dialog allows you to import Jigsaw Datafiles. We have created a simple proprietary xml file format for Jigsaw. If you have some specific type of data that you would like to analyze in Jigsaw, one option is to first translate it to Jigsaw's Datafile format. We have included a few sample Jigsaw Datafiles in the distribution such as the documents from the 2007 VAST Symposium contest, all the paper abstracts from InfoVis and VAST papers, a sample of papers from PubMed about breast cancer, and the Bible. We use the folder named `datafiles` to store Jigsaw Datafiles. More about this is presented in the next section and in the Appendix.

When you import a document or a set of documents, you also can choose to perform entity identification on the documents if you would like. This is done via a second dialog box that will pop up. (If you have many files and they are relatively large, entity identification can be time-consuming, so be patient.) Alternately, entity identification can be performed later in the analysis process as well. To learn more about entity identification, see Section 3.

When Jigsaw imports a set of documents, it builds an analysis database for those documents on disk. This is done so that Jigsaw can scale up to large document collections. Note, however, that the first time a set of documents is imported, building this database can be time-consuming, perhaps taking quite a few minutes. Once done, this analysis database is called a Jigsaw *Project*. Subsequent analysis sessions will be much faster to commence though since this Jigsaw project/database simply will be read in from disk. A Jigsaw Project file (.jp) is represented by a file in the `Projects` folder of the system and it encapsulates a set of documents that have been read into Jigsaw along with any entity identification that has been performed on them. Jigsaw also uses the concept of a *Workspace*. Workspaces include all the information of a Project, but they also include multiple Views that may have been active during an investigation. Jigsaw Workspaces (.jws) are encapsulated by a file stored in the `workspaces` folder. The *File* menu in the Jigsaw control panel includes commands for opening and saving Projects and Workspaces.

1.3 **Displaying Views**

To begin analysis, you likely want to start with a set of views. Go to the *Views* menu and choose whichever ones you want. Note that you can create multiple instances of any view type. We highly recommend having at least one Document View open all the time. Note that almost all Jigsaw views begin empty (the Document Cluster View is an exception). This is normal. You must perform a search query or do some command in a view in order to look for entities and/or documents and to begin to populate the views.

1.4 **Start analysis and exploration**

To begin exploration, you can enter a search term in the Control Panel. Jigsaw will look for any identified entities containing that text and will display an appropriate representation in each of the views present. This is the default *Entities* search mode which is initially selected. The *Documents* search mode is useful when you want to search for a plain word (e.g., dog, car) that is not necessarily an entity. Jigsaw acts more like a simple search engine for this, bringing up the documents that include the search string.

In general, there are three ways to populate and add information to the views:

1. Issue search queries from the Jigsaw control panel.
2. Right click on an entity or document and issue the *Show* command. This pushes that item out to be displayed in all the other views that are listening.
3. Double-click on an entity or document which issues an *Expand* command and shows connected items in that view and other listening ones.

1.5 **Saving a session**

You can save an analysis session already underway by saving it as either a Project or a Workspace. These commands are available under the *File* menu. See the next section for additional details about projects and workspaces.

Important Notes: There are some performance issues when running Jigsaw with anti-virus software (e.g. McAfee or Norton/Symantec) when the auto-protection feature is enabled. This feature scans the Jigsaw database files each time they are accessed, resulting in extreme slowness. To improve performance, you need to exclude the Jigsaw/projects folder from being scanned during execution.
Below are procedures for McAfee VirusScan 8.7i that you can use to improve Jigsaw performance while still allowing your system to have virus checking enabled:

1. Open McAfee *On-Access Scan Properties*
2. Click on *All Processes* and then the *Exclusions* tab
3. Click the *Exclusions* button and then the *Add* button
4. Use the *Browse* button to locate the Jigsaw/projects folder and click *OK*
5. Choose *OK* again
6. Close McAfee VirusScan

Follow similar procedures for other anti-virus tools.

Similar performance issues can occur when using the indexing functionality in Windows Vista or Windows 7/8. You can exclude the Jigsaw database from indexing using the following steps:

1. Enter the phrase “Indexing Options” in the search box on the Start menu
2. Click the *Modify* button
3. In the *Change selected locations* panel at the top, exclude the Jigsaw/projects folder
4. Click the *OK* button and click the *Close* button

2. Importing and Saving Documents

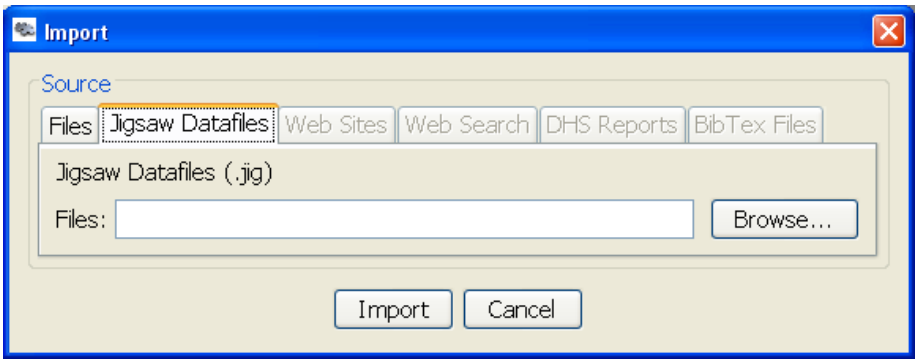
2.1 Importing Documents

Jigsaw can import a variety of types of text files. Presently, it can read in ascii or comma-separated value (.csv), Unicode text (.txt), Adobe Acrobat (.pdf), Microsoft Word (.doc), and HyperText Markup Language (.htm or .html) files. Plain ascii or Unicode text files are the most reliable type of file to import, so whenever possible we recommend that you use text files or transform your documents into text files if possible. Jigsaw also does a reasonable job importing "simple" Acrobat and Word files made up primarily of plain text. However, if an Acrobat or Word file has complex formatting, images, etc., then the import process is much less reliable. Jigsaw can only read old style Word files (.doc), not new .docx files. Also, beware of Acrobat files with security or password protection. When possible, we recommend transforming Acrobat or Word files to plain text for more reliable import. We have scripts and tools that can automate this transformation for large numbers of files. Ask us about this if you are in that situation.

For html files, Jigsaw attempts to discard tags and only read in the actual text content, a process which is not perfect. In general, the process of importing html documents in Jigsaw still needs to be improved so do that with caution.

Note that Jigsaw views source documents as all textual content. In general, any text within the file is viewed as the body of the document. There are two exceptions to this, however. If Jigsaw finds the string "Date:" or "Source:" followed by some other text on a line within the top five lines of a file, then it interprets that as a special meta-data line and it uses the trailing string as the special <DocDate> or <DocSource> fields for the document, respectively.

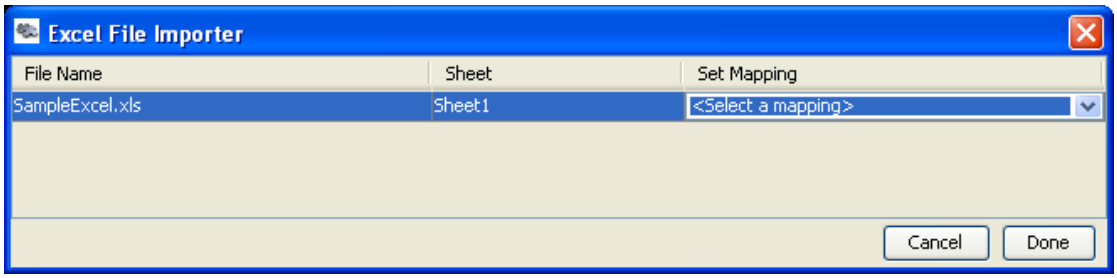
To read in multiple files at once, simply select multiple files in the File Chooser dialog box using the shift- or control- mouse selection operation for your particular operating system.



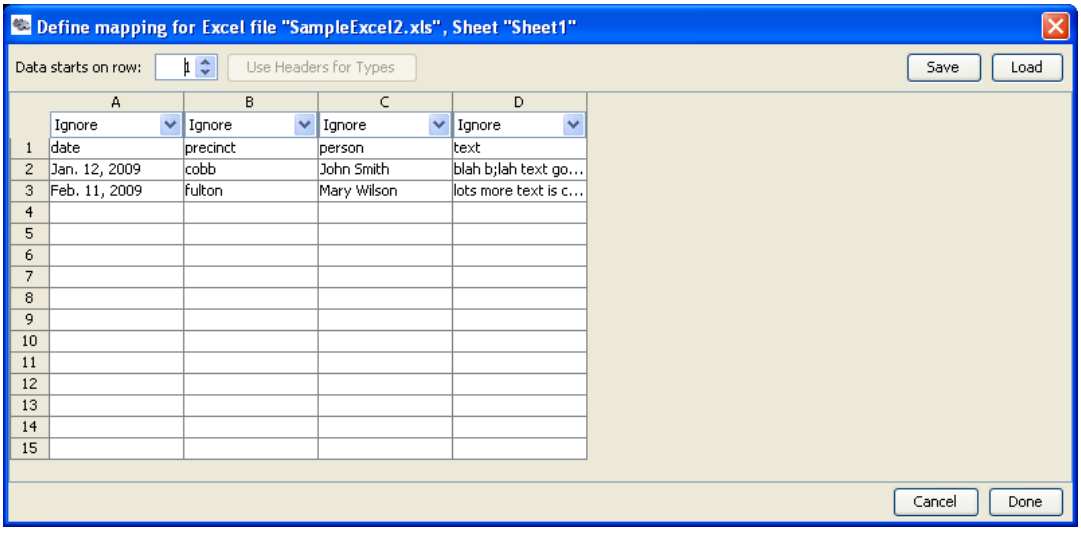
Importing MS Excel and csv Files

Jigsaw also can import Microsoft Excel files saved as .xls or .csv files. We *strongly* recommend using .csv files whenever possible as this is much more reliable than reading the formatted Excel files. Also, it's really simple to generate .csv files from your .xls or .xlsx files.

Because the primary unit of analysis in Jigsaw is a document, you inevitably may wonder how this is handled with these types of files. In general, Jigsaw considers each row of a sheet as a separate document. The columns in a spreadsheet can specify attributes such as the ID, date, or body text of the document (row), or they can be a type of entity. It is your responsibility to set up the mapping from columns to the relevant attributes. When you initially import a spreadsheet file or files, you will be presented with a dialog box that allows you to either define the mapping or load a pre-existing mapping. You should use the pull-down menu in the right-hand column below and choose "New" to create a new mapping or choose one of the existing mappings already there.



When defining a mapping, you will see a dialog box like the one below. It allows you to define the attribute specified in each column by selecting the pull-down menu above that column. The menu contains items for the Document ID, date, text, and for common entity types such as person, place, and organization. This menu also allows you to create a new type of entity to be specified in a column. At the top of the dialog box, you can specify the row in which the actual data begins, thus ignoring some header rows.



Some important points about spreadsheet import:

- Jigsaw can only read old style Excel files (.xls), not the newer (.xlsx) files. In either case, we recommend using .csv files instead.

- When importing Excel (xls) documents, make sure none of the cell contents are specified by formulae. The Jigsaw import process will not work in that case. To fix this, choose all the cells from the sheet, make a copy, and then paste the cells *as values*. The software that we use to read Excel sheets needs to have simple values, not the formulae that are sometimes stored in Excel files. Of course, this isn't a problem if you use .csv files.
- If you create a new entity type in a column, that entity type's name can contain only letters and numbers, and it must start with a letter. No other characters are allowed.
- If some of your cells are empty, then the results may be unpredictable. Most of the time we believe that they simply will be skipped and it will work "correctly", but to make sure of success, try to have contents for all cells.
- If possible, try to specify the Document ID and the Document text attributes. Even if you choose some simple text column to serve as the Document text, this will be helpful. You might even make a new column in your spreadsheet that is the union of a variety of other columns.
- If Jigsaw finds duplicated Document IDs in a sheet being read, the first one will be used, following ones will be skipped.

Jigsaw Datafiles

We have created a proprietary file format for storing collections of documents that uses xml. In addition to the text contents of a document, this format can contain meta-information about the document such as an ID and a date, and it can hold a list of identified entities for each document. We call these proprietary files Jigsaw Datafiles (.jig). By convention, we store these in the `datafiles` folder of the distribution. We have included a number of samples there for you to examine. **Important**: Make sure to choose the second tab in the *Import* dialog box to read in Jigsaw Datafiles.

If you have your own data perhaps in some xml format, in a database, or in another format, it's not too difficult for you to translate this into Jigsaw's Datafile format. Examine the Appendix of this tutorial for more information about that and for instructions on how to work with your own data. Trust us -- It's really not too bad. We have done this to convert other xml files into Jigsaw's format and to scrape web pages and make Jigsaw Datafiles from them. Remember that this is xml, however, so you cannot have characters such as &, %, <, or > in your text. The Appendix also has more information about this.

The first line of a Jigsaw Datafile can be a filetype specification (Unicode utf-8, for example). Jigsaw will read this specification and interpret the file correctly.

Note that if you create your own Jigsaw Datafile and you try to import it and the process fails or hangs, then you likely have a syntax error in the file such as an illegal character, a missing bracket, a mismatched open/close tag, etc. View the small import indicator dialog that is shown and it should give you a clue as to where the syntax error is. For example, if it hangs while importing the 17th document, then there is likely a syntax error in it or the one after it in your file.

As another option, if you have your own specific data file format and you are not sure how to put this into Jigsaw, please get in touch with us and we can possibly write an importer for that file format or a translator from that into Jigsaw's Datafile format.

Note that if you have imported documents from text files, Word files, spreadsheets, etc., and you would like to see them in Jigsaw Datafile format, the *File* menu also has an *Export* command for writing out the current project as a Jigsaw Datafile.

Jigsaw Control Panel Information

When Jigsaw has successfully imported a set of documents, it shows information about the documents in the Control Panel window. It will show the number of documents that have been read in successfully, and the different types of entities with a count of how many unique values each entity has. Jigsaw assigns a unique color to each entity type that is shown across all the different views in the system.

2.2 Jigsaw Projects and Workspaces

When a set of documents have been successfully read in and entity identification potentially performed, this set of information is called a Jigsaw Project. A Jigsaw Project file (.jp) is represented by a file in the `projects` folder of the system and it encapsulates a set of documents that have been read into Jigsaw along with any entity identification that has been performed on them. You can save projects and then reopen them on subsequent runs of the Jigsaw system. Simply use the *Save Project/Save Project As* and *Open Project* commands from the *File* menu in the Jigsaw Control Panel.

As with any prototype system, we recommend that you save your project manually to be safe. Note that once a project has been saved, you only need to re-save it if you modify the entities within the documents. Also, for reasons that are too complex to be explained here, the *Save As* command will take a longer time to execute than the *Save* command.

When Jigsaw imports a set of documents, it builds an analysis database for those documents on disk. This is done so that Jigsaw can scale up to larger document collections. Note, however, that the first time a set of documents is imported, building this database can be time-consuming, perhaps taking quite a few minutes. Subsequent analysis sessions will be much faster to commence though since this Jigsaw project/database simply will be read in from disk. These databases are stored in a folder named `DBs` under the `projects` folder. Many files may be stored there and they may use a significant amount of disk space. That is normal.

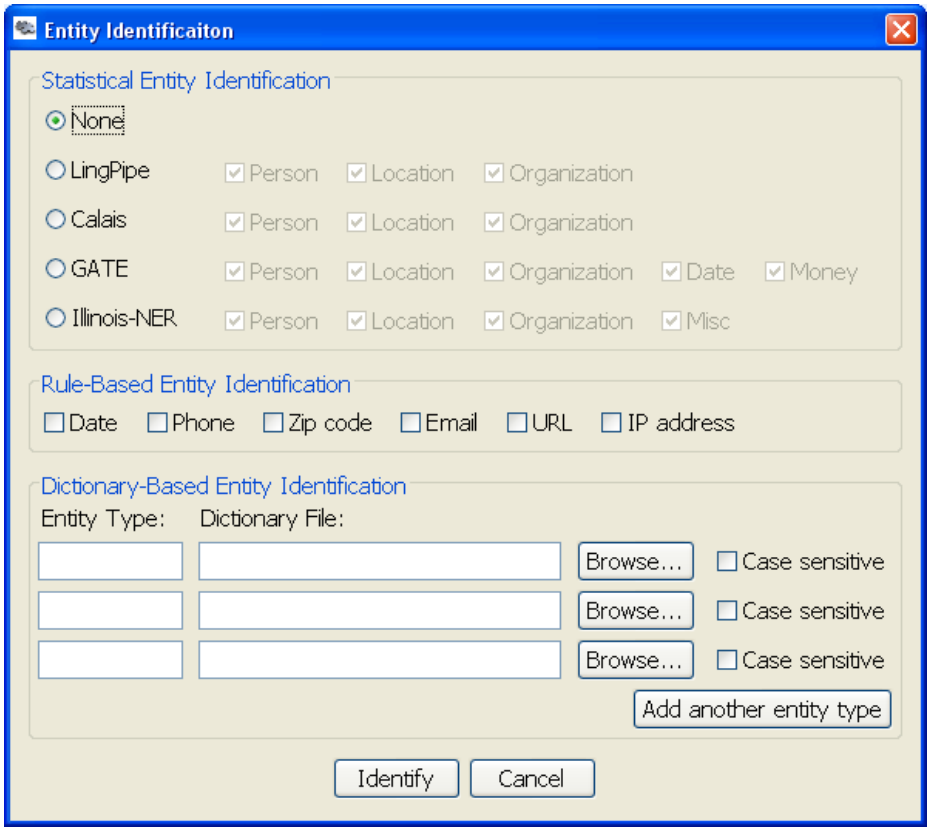
In addition to a collection of analyzed documents, you can save the active set of Jigsaw views as well. This combined collection of documents and views is called a Workspace in Jigsaw (.jws). Just as done for Projects, the *File* menu in the Jigsaw Control Panel contains *Save Workspace* and *Open Workspace* commands for manipulating Workspaces.

3. Identifying and Working with Entities

The Jigsaw Control Panel contains an *Entities* menu that includes operations for the different entity processes described below. Also, when new documents are imported, Jigsaw will prompt the user about performing entity identification.

3.1 Entity Identification

When importing text files or spreadsheets, you can choose to have the system automatically identify entities. Presently, Jigsaw provides three main mechanisms to identify entities in documents. First, it includes third party software libraries to do automated (statistical) entity identification. Second, it includes the capability to do some basic pattern matching of text to identify entity types such as dates, phone numbers, zip codes, email addresses, URLs, and IP addresses. Third, it allows you to provide an entity type (name) and a list of the values of that entity type. Below we describe each of these in a little more detail.



For automated entity identification, Jigsaw can apply one of four possible packages. Lingpipe (http://alias-i.com/lingpipe/), GATE (http://gate.ac.uk/), and the Univ. of Illinois entity recognition system LBJ (http://l2r.cs.uiuc.edu/~cogcomp/asoftware.php?skey=FLBJNE) are included with the distribution, so the entity identification process will be done local to your computer in these cases. The OpenCalais web service requires that you have an internet connection to use it. When you do this for the first time, Jigsaw will prompt you to get an API key for using the service. You will need to go to their website to retrieve a key, but this only

needs to be done once. Please be aware that your documents are sent out to the OpenCalais server to be processed, so please do not use it for sensitive or private information. All the packages have strengths and weaknesses so we recommend you try each to see which will work best for your documents. We generally use the Illinois NER system and have found it to be quite good in general. Do note, however, that it often seems to crash when trying to read pdf, Word, or html files. The GATE system works quite well too, so you may want to give it a try as well.

Jigsaw also contains functionality that can help you identify particular types of strings such as dates, phone numbers, zip codes, email addresses, URLs, and IP addresses in documents' text. This code does some basic regular expression matching so it is not perfect. For example, a 5-digit number will be identified as a zip code; we do not validate this with all actual zip codes in the United States.

Finally, Jigsaw allows you to create a new entity type and specify all of the valid strings that are the instances of that entity. For instance, you could create a new entity type "Car" and specify a set of possible values such as "Ford", "Chevrolet", "Honda", "Hyundai", etc. To do this, you need to create a text file (.txt) that has each different possible entity value on a different line of the file. (Note that an entity value needs not be just one word; it can have multiple words.) Jigsaw also allows you to specify limited types of regular expressions through this entity definition file as well. To designate that a line of the file is a regular expression, the line must begin with the three characters ">>>". After that, you simply put the letters and numbers that you want along with special pattern characters: &-denotes a letter, #-denotes a number, $-denotes either a letter or number, +-denotes one or more occurrences of the previous symbol, *-denotes zero or more occurrences of the previous symbol. Thus, the string "`&red#*`" tells Jigsaw to find any word starting with a letter, followed by "red" and then followed by zero or more numbers and consider it a valid entity value. The string "`abc&&&`" tells Jigsaw to find any word beginning with "abc" followed by three letters as a valid entity value. An entity definition file can have a mixture of normal (non-regular expression) and regular expression lines throughout.

To then add this new entity type to Jigsaw, you use the bottom region of the entity identification dialog. Simply enter the entity type name to the left and then browse for the text file containing the list of entity values. Note that entity type names (such as "Car" in the example above) are case-sensitive, can only contain letters and numbers, and must start with a letter. In using this third type of identification you can make the matching be either case-sensitive or not.

The *Entities* menu in the Control Panel provides a command for doing additional, subsequent entity identification as well. (*Note*: There is a bug in Jigsaw and sometimes this command does not function properly.) Entity identification can be run at any time in an investigation, not just when documents are initially imported. If you have any views open, however, they will be closed at that point because the data they show may subsequently be incorrect.

3.2 Correcting Erroneous Entity Identifications

The process of automated entity identification is not perfect. Many false positives (identifying entities that really are not entities) and negatives (completely missing some

valid entities) can occur especially in documents with many spelling errors from processes like OCR.

Jigsaw provides the ability to fix incorrect entity identification. In the Document View, you can right-click on an entity and a menu will give you options to remove the entity or change its type. Furthermore, you can mouse-click-drag over words in a document to select them, then right-click on the selected word(s) and you will get a menu that allows you to add the word(s) as an entity. You can choose one of the existing entity types or you can create a new entity type. The identification also can only be relevant to that one particular document or it can apply to all documents in the collection.

The List View also includes a right-click-menu command *Delete* that allows you to correct erroneous entity identification and remove an entity or entities. You can select multiple entities with shift- or control-click in order to remove multiple entities at one time.

New entity types (names) are case-sensitive and cannot contain blank characters or other special characters. The entity type can only include letters and numbers, and they must begin with a letter.

3.3 Manipulating Entities

The *Entities* menu in the Jigsaw Control Panel provides some other operations particularly useful for managing entities.

The *Clean Up Entities* command goes through all documents in the collection and identifies "missing" entities. For example, if "Mary Wilson" has been identified as a Person entity in one document but not in some other document, by issuing this command it will be added as a Person entity in the second document. Essentially, this command makes sure that any entity identified in some document is consistently identified throughout the entire document collection. Be aware that this function also may add more "noise"; if an entity was incorrectly identified in one document, it will be added to all other documents in which it occurs.

The *Remove Singles* command simply removes any entities in the entire document collection that occur only one time. This command essentially allows analysts to remove potential "noise" entities and focus on those that more commonly occur. Note that this command can only be issued before any views have been created.

Note: Because of how they are implemented, both the *Clean Up Entities* and *Remove Singles* commands will run extremely slowly on larger document collections. We hope to fix this in a subsequent version of the system.

The *Entities* menu also includes a *Manage Entities* command that will bring up the entity browser for viewing entity attributes and for manipulating entity aliases (see below section). So, for instance, you could bring up a person entity and you could create a phone number attribute and enter the person's phone number as that attribute's value. Presently, Jigsaw does not incorporate entity attributes into the system extensively, but you can specify them manually through the entity browser.

3.4 Entity Aliasing

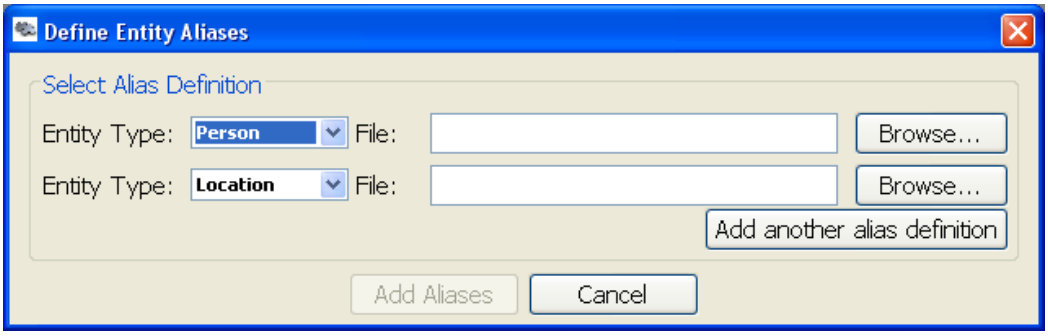
Jigsaw also allows you to create aliases for entities. Suppose that a person's name is spelled three different ways in a document collection, but you know that they are all the same person. Alternately, suppose that a person is using an alias, that is, there is another

name that they go by. Jigsaw allows entities to be aliased in order to handle either of these situations. Entity aliases can either be defined interactively through the views or by specifying an entity alias file that is read in at the start of an investigation.

In order to interactively create an alias, select two or more entities in the List View or the Graph View (remember to use control-click to accumulate selections), and then right-click to invoke a menu that will have the *Make Aliases* command in it. Choose that, and the system will ask which of the entity names should be the main one to use for this alias. Once you have done that, all the other subordinate entities will be removed from views and only this main entity name will be used. That "winning" entity name will be drawn in an italic font to indicate that it has aliases. Upon moving the mouse cursor over such an entity, a pop-up view will arise showing the other aliases.

In the List View, if there are too many entities, this pop-up behavior can be annoying. You can turn this off via a menu command at the top. If you do, the aliases for an entity will only pop up if the F1 key is held down as you mouse over the names.

To specify entity aliases via a file, use the *Add Aliases* command from the *Entities* menu in the Jigsaw control panel. We recommend you do this right after you have imported your documents and run the initial entity identification. When you issue the *Add Aliases* command, the dialog box shown below will appear.



For each different entity type that will have aliases, you need to create a text/csv or a JSON file that has a set of alias specifications in it. Each line of the file will hold a different alias. On that line, use commas to separate the different strings to be aliases. The first alias on the line will be the one shown throughout Jigsaw as the "winner" in an italics font. For example, below are three valid alias definition lines in a txt or csv file:

```
"Brown University", "Brown Univ.", Brown, Bruno
CIA, "Central Intelligence Agency", "C.I.A."
```

For a JSON file, the format looks like

```
{
   "Brown University" : ["Brown Univ.", Brown, Bruno],
   "CIA" : ["Central Intelligence Agency", "C.I.A."]
}
```

The *Manage Entities* command also invokes the Entity Manager window that allows you to further modify aliases.

4. Exploring and Analyzing a Document Collection

Once you have imported a document collection, you are ready to explore, investigate, and analyze the documents and their entities. In all likelihood, you want to create a number of different views to show the documents and entities. Remember that you can have any number of views of any of the existing view types present.

4.1 General Tips

- Views show entity-document and entity-entity connections. A document and an entity are connected if the entity appears in the document. Two entities are considered to be connected if they appear in at least one document together. As the number of documents in which they appear together increases, so does the quantitative connection strength.
- Any view can be cloned (copied) at any time using the *Edit* menu's *Clone* command.
- The *Edit* menu's *Clear* command in a View is useful if the view gets too "noisy" and you just want to blank it out again.
- A single mouse click on an item (document or entity) selects that item. All the other visible items then update their appearance to show how they related to that selected item. A double-click on an item expands the item – typically this shows connected items to it. User mouse actions such as selections and expansions also are transmitted to other active views which update their representation appropriately too.
- You can turn off/on event listening in each view by clicking on the little satellite dish in the upper right. Turning off listening essentially freezes the view, that is, user actions such as clicks and double-clicks in other views will not affect this view. This capability is very useful to lock a view at an interesting state. Note that frozen views also are not affected by the *Clear All Views* command in the *Views* menu.
- To push an item (entity or document) out to all other open and listening (non-frozen) views, right-click on the item and use the *Show* command. Similarly, doing a search on a string through the Control Panel will push out all entities containing that string to all the active (listening) views.
- To examine a document or the set of documents containing an entity in an empty new Document View, right-click on the item and use the *Show in new Document View* command.

4.2 Search Tips

There are two search/query modes available through the two checkboxes under the query entry region in the Jigsaw Control Panel window. In *Entities* mode, Jigsaw will seek out documents containing the words from the query string in already identified entities within those documents. That is, only documents (and entities) will be retrieved that have the search terms in existing entities in those documents. This is the default mode. In *Documents* mode, which is accessed by selecting the *Documents* checkbox, Jigsaw simply retrieves documents that contain words from the search query somewhere in the document text.

Note that different kinds of Boolean searches can be performed in Jigsaw. (For those in the know, we use Lucene to perform search.) When your search query has multiple words such as

```
John Mary Bill
```

you will be doing an "or"-based search, that is, finding documents that mention one or more of those words. You can also do other Boolean operations such as

```
"john smith" AND mary
```

which searches for documents having both "john smith" and "mary" in them. For more about the different types of searches that can be performed, see http://lucene.apache.org/java/3_0_0/queryparsersyntax.html.

4.3 View-specific Use Tips

The sections below briefly describe some of the utility, commands, and capabilities of the different views in Jigsaw. For more specific help and use tips, please see the video tutorial for each view at `http://www.gvu.gatech.edu/ii/jigsaw/tutorial`.

Note that each view has its own menus at the top that provide useful operations for that view. For example, some of the views have filtering operations that allow you to limit what is shown. All views have *Change Title*, *Clear*, and *Clone* operations and the ability to bookmark the view.

Document View

The Document View is the core view in Jigsaw for reading document contents. The list on the lower left holds a set of documents that have been loaded into this view. Documents are placed there in response to control panel search queries, by *Show* commands from other views, or by *Expand* commands issued in other views. Additionally, the *Add All* button in the lower left will bring all documents in the collection into the view. Be careful about using this command with extremely large document collections.

Click on any document name to select it and show its text in the focus area to the right. The number by the document ID is how often a document has been viewed. All of the documents listed in this view that have the little clouds to the left selected (darkened) are participating in the word cloud at the top, which shows the key words used throughout that set of documents. You can click on the clouds to toggle whether or not the corresponding document participates in the word cloud.

In the region above the actual document text is the "document summary," the one sentence from the document that Jigsaw has selected to most exemplify what the document is about. This can be useful for fast triage of multiple long documents.

Within the document focus region, the body text for the document is shown at the top, and below it are listed any affiliated entities that do not occur in the document text. Entities are colored in a pastel shade of their default color. Clicking on an entity selects it. You can perform manual entity identification by making a mouse drag selection of a word or words which selects them, then do a right mouse button-click on the selection to get a menu that allows you to add this as a new entity. Similarly, you can right-click on an already existing entity to access commands for removing it as an entity or changing its entity type.

Note that as documents get larger and larger, they tend to load much more slowly in the Document View.

The Document View has a menu command at the top for printing out to a file all of the loaded documents within it.

List View

We find the List View to be the most powerful and useful view in Jigsaw. It provides very easy browsing, selection, filtering, and exploration of all the entities and documents in the collection being analyzed.

The view begins showing two columns, but you can add/remove lists (columns) via commands from the *Lists* menu in the view so that you can fill out a wide view with as many lists as you want. The view will scroll horizontally if there is not enough room.

Each column holds entities of a particular type – the type can be changed through the menu at the top of each list. The same entity type can be put into different columns too. You add entities to lists by doing search queries or issuing *Show* operations from other views. Additionally, the *Add All* button above a list brings all the entities of that type from the entire document collection into a list. This is often useful early in analysis to examine all the different entities that have been identified. Be careful with very large document collections with many, many entities of a particular type, however. That may generate a very long scrolling list.

The bar to an entity's left is a frequency counter across the entire document collection. By moving the mouse pointer over this small bar you can find the exact number of documents in the collection in which that entity appears.

The buttons and menus above a column control how that particular list appears. The first three buttons sort the list in different ways: 1) alphabetically, 2) by frequency of appearance in the entire collection, or 3) by connection strength to the selected item(s). Other buttons control the alignment of entities and allow you to clear a list.

Clicking on an entity selects it; shift-click and control-click allow multiple entities to be selected. Selected entities are drawn in yellow. Entities connected to the selected entities are drawn in orange with darker shades indicating stronger connections. Unconnected entities are drawn with a white background. When multiple entities are selected, two buttons in the upper right control whether entity connections are shown via or'ing the selected entities or and'ing the selected entities. For example, in *and* mode, connected entities (those shown in orange) must co-occur in some document(s) with all the selected entities.

A right-mouse-button click on a selected entity or entities provides a menu with a number of useful operations including *Show*, *Hide*, *Expand*, and *Delete*.

The List View has a menu command at the top for printing out to a file all of the currently visible entities in the lists. If you want just the connected entities, sort the list by connection strength and then just remember the last connected entity and edit the output file manually. (This is currently broken in the distribution.)

Graph View

The Graph View provides a node-link graph representation of the document and entity collection. To have documents or entities appear in this view initially, you must search for them in the Control Panel or issue *Show* commands from other views. Documents are represented by white rectangles and entities are represented by colored (by type) circles. A line from an entity to a document means that the entity appears in that document.

A single click selects an item and double click expands/compresses it. You can click and drag on the background to do a rectangular rubber-band selection of multiple items at once. Shift-drag-click does a circular rubber-band instead. Expanding an entity (double-click) shows all the documents in which it resides, and expanding a document shows all the entities within it. You can manually move documents and entities by clicking and dragging them. The plus sign indicates that not all connections (entity-document) are presently shown and you can expand it to see more. Right-click gives a menu of helpful operations. The *Flow Select* command is useful for broadening the active selection out by one jump. You can then issue a subsequent menu command such as *Expand*, *Remove* or *Invert Selection* and it applies to all of the selected items.

The *Circular Layout* button/command at the top left is very useful. It changes the layout of documents so that all are equally spaced around a virtual circle in the window. Entities appearing in only one document are drawn near it, but outside the circle. Entities appearing within multiple documents shown in the view are drawn inside the circle. As the number of document connections increase, the entity is drawn closer and closer to the center.

The cursor button near the upper right toggles between Selection mode (default to start) and Zoom/Pan mode. In Zoom/Pan mode, clicking down with the left mouse button and dragging does a pan operation. Clicking down with the right mouse button and dragging does zoom out (mouse movements up) and zoom in (mouse movements down).

The default background color for the Graph View is black unlike almost all other Jigsaw views. It can be changed to white through a command in the *Options* menu at the top. This menu contains other commands for customizing how the graph is drawn.

Calendar View

This view shows documents and entities in the context of dates relevant to them, that is, in the context of the formal Document Date or any dates mentioned within a document. The main portion of the view shows a calendar. Within that region, small gray diamonds represent documents and colored diamonds represent entities. The diamonds are drawn on the dates relevant to the entity/document that they represent, ie, the dates from the documents in which they appear. When an entity is found using search or the *Show* command, it is added to the upper left region of the view, but it is not initially placed in the calendar region. By clicking on the entity name in the upper left region, you can toggle whether it should appear in the calendar region of the view. Initially, the default entity type color is used for each entity, but by clicking on the small color square, you can change this color for a particular entity (perhaps to make it stand out). By clicking on an entity type name in this area, you toggle whether all of the entities of that type below are shown or not. When you move the mouse pointer over a document diamond, the entities in that document are shown in the lower left. The *Show All*, *Clear All*, and *Filter* menus at the top are useful to control which types of entities are shown.

The controls in the upper left corner of the view are very important. You can have entities and documents appear at the actual DocDate of a document, or you can also have entities and documents appear on all of the different dates mentioned within a document, or you can have both. You also can control the range of years shown in the view. The two modes within the Granularity menu control whether individual days are shown in the calendar (larger, likely requires scrolling) or only months are shown (more compact).

Document Cluster View

This view provides a quick overview of the entire document collection by representing each document in the collection as a small rectangle icon in the window. Initially, the documents start in one big pile, but they can be manually moved or clustered by operations in the view. When a search query from the Control Panel is done, the search term will be added to the upper left region of this view. Clicking on that term will then color all the documents that contain that particular entity. Clicking on *Group by Filters* button then will segregate the documents into clusters by color.

In addition, if you have performed a document clustering computation (through the *Tools* menu in the Jigsaw Control Panel), the documents can be laid out in clusters determined by either the full text of the documents or the entities within documents. Use the pop-up combo chooser to select which clustering you want to display. Through the Tools menu in the Jigsaw Control Panel is a command to create and name new clusterings (varying in number of clusters and whether they are text- or entity-based). This area also contains a slider that allows you to control the words selected to describe what each cluster is about. These terms can range from being very frequent, common ones (slider to the left) or very unique words only to that cluster (slider to the right). You can right click on any of the descriptive words for a cluster which will bring up a menu that will allow you to replace the word and do other commands.

Note that by clicking on the header for a cluster, you can select all the documents within that cluster and perform an operation on all or simply move all together.

You can also highlight (in a surrounding yellow glow) all the documents that you have read so far (been brought to the front in some Document View) by clicking on the *Highlight Viewed Documents* button in the upper left. This is useful for keeping track of what you have examined so far in an investigation.

The bottom-left region shows a Windows Explorer style view of the hierarchical clustering of documents. You can click on documents in the list to and their corresponding icon will be selected. Clicking the plus-sign for a particular document also shows all the entities in it.

Documents can be manually moved and grouped by clicking and dragging as well. Clicking in the background and dragging performs a rubber-band selection so that multiple documents can be chosen and moved together. A regular click-drag does a rectangular selection and a shift-click-drag does an oval/circular selection.

The Document Cluster View has a menu command at the top for printing out to a file all of the different clusters and which documents are inside each cluster (currently broken in the distribution).

The default background color for the Document Cluster View is black unlike almost all other Jigsaw views. It can be changed to white through a command in the *Options* menu at the top.

WordTree View

This view is a version of the WordTree visualization introduced by IBM through the Many Eyes visualization site and their 2008 IEEE InfoVis paper. Here, the WordTree applies to all documents in the collection. This view helps you understand the context of different words in the collection.

When you enter a term in the upper text entry region, the system will show all the trailing words/context that follow it in some document. You can constrain the view to compress all the strings to fit in the window or you can allow it to show more via scrolling. By right clicking on a string in the view, you can bring up a menu that allows you to add those documents containing the string to a Document View window.

Document Grid View

This view is useful for seeing a sorted and shaded list of all the documents in the collection where the order and shading can communicate different metrics about the

documents. The view begins empty but documents can be added via *Show* operations in other views, search queries, or the *Add All* button in this view. Each document is represented as a small rectangle within the view. The documents are sorted from the top-left to the bottom-right by row. You can apply different metrics to control this order and the shading of each document's rectangle. Mousing over a document rectangle shows its Document ID and the value for the metric used to control sorted order. Presently, only a number of different metrics are available: the size of a document, the number of entities in a document, the document date, the document's sentiment, and the documents' similarity to a selected document. By selecting the checkbox in the upper left, you can make the documents organized by cluster (if that has been computed) and then ordered and sorted appropriately within those clusters.

The Document Grid View has a menu command at the top for printing out to a file all of the different documents in the view in the order in which they appear and with a metric for each.

Circular Graph View

This view is similar to the Graph View, but no documents are presented. Instead, entities that are "connected" (co-occur in some document) are shown with lines connecting them. The view uses a simple graph layout algorithm: all entities are plotted on the circumference of a circle with entities of the same types grouped together. You must select an entity or entities by clicking on them to see the connecting lines. Use control-click to select multiple items at a time.

Scatterplot View

This view is useful for seeing entity-entity connection (co-occurrence) in documents within the collection. You can place different entity types on the x and y axes through the menus at the top-left and bottom-right. Particular entities then are added along the axes through search queries or *Show* commands from other views. A diamond in the center region then represents a document that contains the horizontal-vertical combination of the two corresponding entities. By double-clicking on an entity, you expand it and its connected entities are added to the axes as well. Each axis has zooming sliders to narrow down the viewed region(s) of entities.

Timeline View

This view shows horizontal timelines onto which documents are placed. A tower of small colored bars (entities) represents a document. Each tower (document) is drawn at a position along a timeline corresponding to its document date. You can use the mouse to focus on a particular segment of a timeline. Simply perform a drag selection of a horizontal subregion of a timeline. That smaller temporal selected region then will be drawn in a new timeline above. Any documents falling within the selected region will be shown in this new timeline as well.

4.4 Automated Computational Analysis

Jigsaw provides a number of different automated computational analyses that can help you explore the document collection. It provides four important capabilities: document summarization, document similarity, document clustering, and sentiment analysis.

To employ these analyses, you must first instruct Jigsaw to calculate them. To do this, choose the appropriate command(s) from the Tools menu in the Jigsaw control panel. If you want to employ these analyses, we strongly recommend that you calculate them immediately after importing your documents and performing entity identification. The "Compute All" command from the Tools menu will perform all of these analyses and when it completes, they will all be available for use. By default it uses clusters of size 20. Alternately, you can compute each of the analysis measures by itself. When you do this for the document clustering, for example, you are presented with more control options, ie, how many clusters and whether the clustering is text- or entity-based. (Note that if not enough documents or entities are present, Jigsaw may create a smaller number of clusters than what was requested.) Whenever you subsequently save your analysis in a Jigsaw Project, all the analyses will be there for the next time you invoke Jigsaw. Note that when you perform the computational analyses, Jigsaw blocks and you cannot perform any other operations. The analyses can take a significant amount of time too. For a document collection of five thousand documents or for larger documents, the analyses may take *hours*. In a situation like this, we recommend that you start the analyses and then do something else in the interim, maybe even run the analyses overnight and return to investigation the next day.

Below we describe each of the analyses and how Jigsaw presents it.

Document Summarization

Document summarization is integrated in different ways in Jigsaw. The Document View shows a word cloud (at the top) of selected documents loaded in the view. The word cloud helps you to quickly understand themes and concepts within the documents by presenting the most frequent words across the selected documents. Jigsaw removes frequent, simple words but does not combine words like "make", "makes", and "making" (stemming) in order to be able to highlight identified entities in the word cloud. The number of words shown can be adjusted interactively with the slider above the cloud. Additionally, the Document View provides a one sentence summary (most significant sentence) of the displayed document. This one sentence summary of a document is available in all other Jigsaw views as well. It can be displayed through a tooltip wherever a document is presented as a symbol or its name. The Document Cluster View also provides keyword summaries for the clusters.

Document Similarity

In Jigsaw, document similarity can be measured relative to complete document text or just to the entities connected to a document. These different similarity measures are of particular interest for semi-structured document collections, such as publications, in which metadata-related entities (e.g. authors or conferences) are not mentioned in the actual document text. The Document Grid View can provide an overview of all the documents' similarity (compared to a selected document) via the order and color of the documents in the grid representation. To do this, click on a document to select it and then

invoke the right menu and choose the command to make it as the basis for similarity. Then go to the upper right and make the order and/or the shading of documents in the grid be based on similarity. In all other views, the five most similar documents can be retrieved with a right mouse button command on a document representation. Note that we have found that the entity-based similarity computation sometimes crashes if some of the documents have a small number of (or no) entities.

Document Clustering by Theme or Topics

Jigsaw also can group similar documents together. Like the calculation for document similarity, document clustering also can be based on either the document text or on the entities connected to a document. Computed clusters are shown in the Document Cluster View or the Document Grid View. Within the Cluster View, there is a chooser for selecting which clustering is to be shown in the view. Each cluster is labeled by three words/terms that describe some of the main concepts within the cluster. Within the Grid View, select the option in the upper left to organize documents within the grid by cluster.

Document Sentiment Analysis

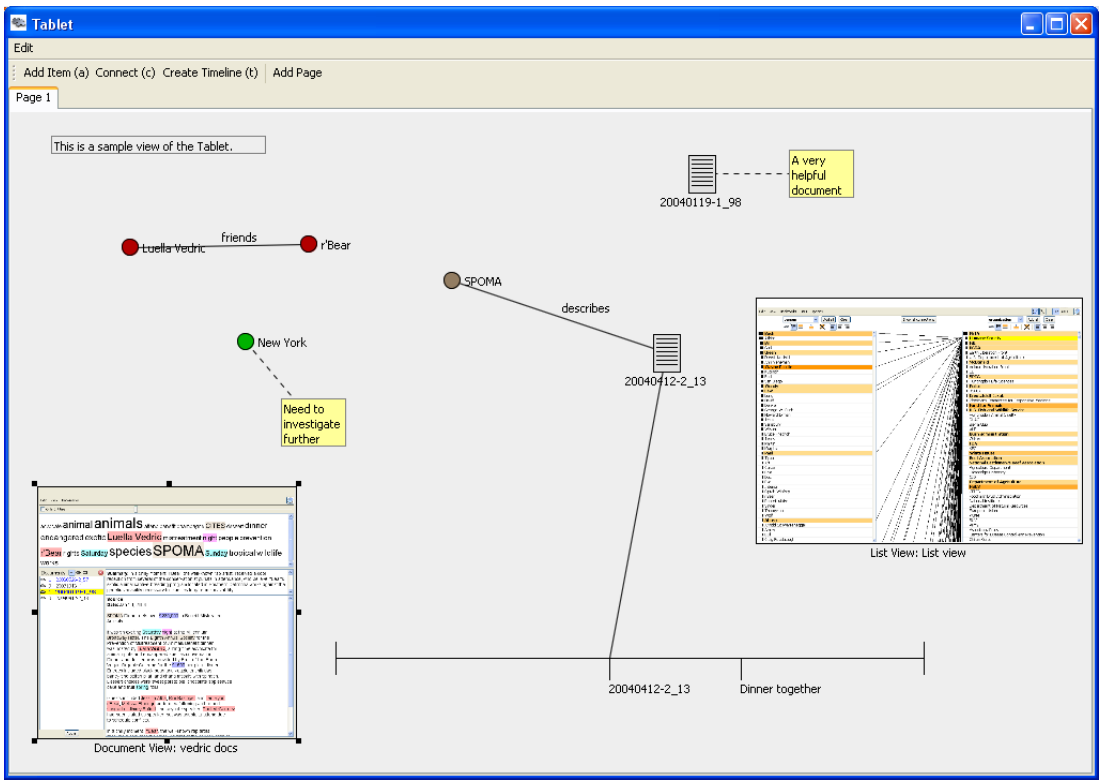
A document's sentiment is its general tone or mood – is it positive and upbeat or is it negative and angry? Metrics about a document's sentiment, subjectivity, and polarity can be displayed in the Document Grid View. Choose the appropriate metric from the menu selections in the upper right. One metric can be represented by the order of the documents, and a second metric (or the first metric again) can be encoded by the document color. To calculate the sentiment of a document, we use lists of "positive" and "negative" words and count the number of occurrences in each document. Jigsaw represents positive documents in blue (more positive is indicated by darker blue) and negative documents in red. You can use your own set of words to determine the positive or negative sentiment of the document as well. Within the *Tools* menu is a command to alter the sentiment dictionary. To do this, you simply create text (.txt) files with one word per line. The command then allows you to either replace Jigsaw's own set of sentiment words with your own or to augment Jigsaw's set of words with yours. Note that these two sets of words need not necessarily be related to sentiment also. You might, for example, create one set of words related to baseball and one set related to football and then the sentiment analysis view in the Document Grid can show whether a document is more baseball-oriented or football-oriented.

Recommending Related Entities for Further Investigation

THIS FEATURE IS CURRENTLY BROKEN. Jigsaw provides another type of analysis within the Graph View. You can select multiple entities (via click then control-click) and then press the right mouse button to get the context menu and choose the command *Recommend Related*. Jigsaw will search for entities that are within two document jumps of all of the selected entities and will generate a window showing all those entities. You can click on any related entity in that view and the path to it from each of the selected items is shown in the bottom region. In that region, you can right-click on a path and issue the *Show* command and then that path (documents and entities) will be drawn in the Graph Views visible at that time.

4.5 Gathering Evidence with the Tablet

Jigsaw provides a window called the *Tablet* that can help an investigator organize his or her thoughts, take notes, gather evidence, develop hypotheses, and so on. Below is a picture of the Tablet with some simple information inside.



An analyst can add entities and documents to the Tablet through right-menu commands in the other system views. Simply perform a right mouse menu click on an item and then choose the "Add to Tablet" operation. Entities are shown as small circles in their appropriate entity type color and documents are small rectangles. Notes (in pastel yellow) can be attached to entities and documents in the Tablet via a right menu command, or notes can simply be placed anywhere in the window by clicking and typing.

The analyst can manipulate objects in the Tablet via the usual Cut (ctrl-x or cmd-x), Copy (ctrl-c or cmd-c), Paste (ctrl-v or cmd-v), and Delete (delete key) commands.

The *Connect* command allows you to connect any two items with a line.

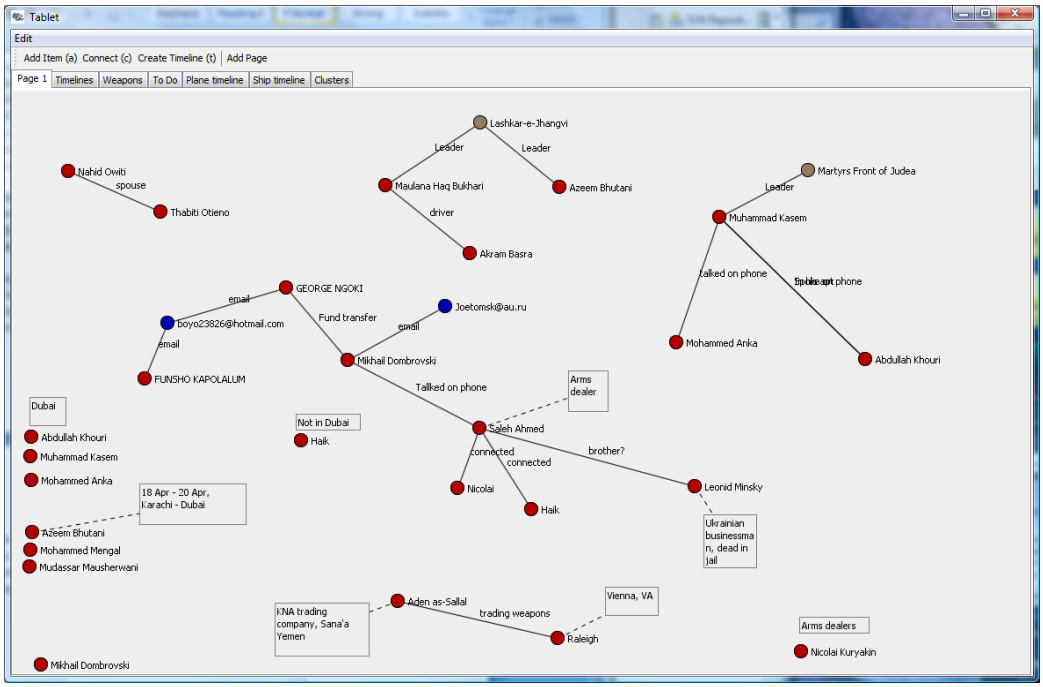
The Tablet also supports the creation of timelines (an example is shown toward the bottom here). To do so, select the *Create Timeline* operation at the top then click down in the window to start one endpoint of the timeline and drag to a position for the other endpoint and then release the mouse button. Events can be explicitly added to the timeline (right mouse click on the timeline and choose the *Add Event* operation), or other items in the window can be connected to the timeline (just drag and drop the item onto the timeline).

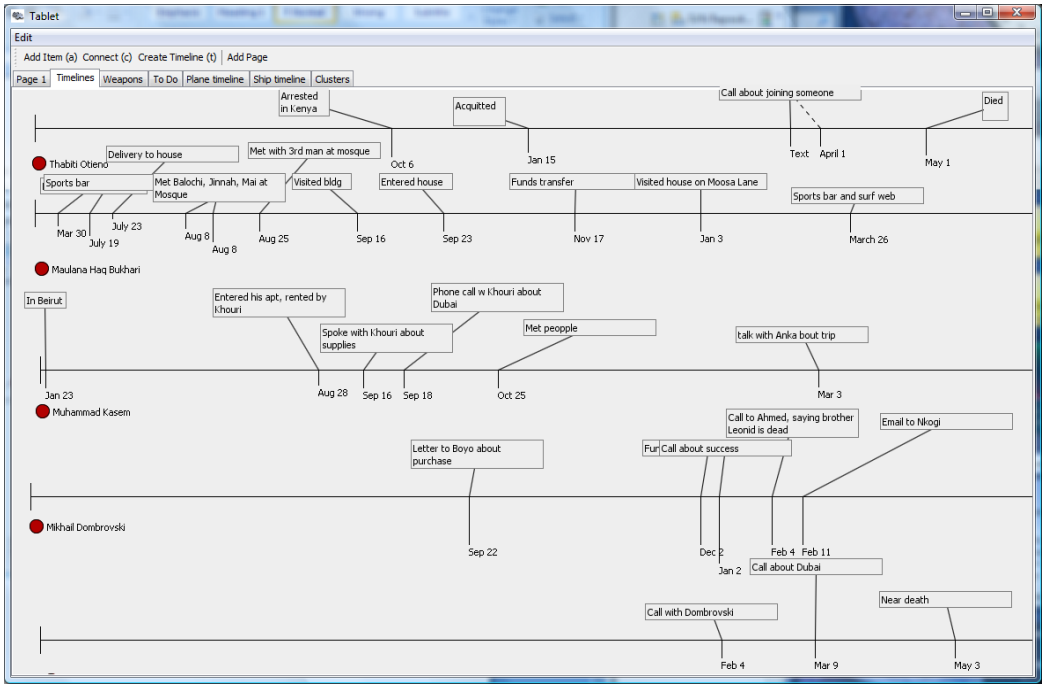
Additionally, bookmarks of Jigsaw views can be added to the Tablet window. Here, you see a Document View bookmark to the lower left and a List View bookmark to the

right. You can add a view bookmark by invoking the "Add as Bookmark to Tablet" command from the *Bookmark* menu in any view.

The final operation at the top, *Add Page*, allows you to add new pages/tabs to the Tablet for your analysis. Below is shown a Tablet with multiple pages, two of which are illustrated. The first shows how you can construct a social network-style diagram from your analysis and the second shows a timeline-focused analysis display.

The Tablet contains a command for exporting the current tab to a PNG image file.





5. Additions to Distribution

Release 0.53 (January 2014)

- Added support to import Unicode documents and Jigsaw datafiles
- Added ability to specify entity aliases externally via a file
- Added "Show in new Document View" command to right mouse button menu
- Added capability to export Tablet as a PNG image
- Modified entity identification dialog to allow more than three new entity types to be specified
- Fixed a bug in entity-based document clustering
- Fixed issue with tree representation in the Calendar and Cluster View on Linux
- Added horizontal scrolling support in List View
- Fixed a bug when importing csv files with empty rows
- Fixed tooltip and changed icon for "Sort by document date" in Document View

Release 0.52 (July 2013)

- Fixed a bug in how dates were compared
- Lists with numeric values are now sorted in numeric order in List View
- Bugs in document summarization fixed. Summary is now always one sentence and its tooltip is not allowed to be too large.
- Fixed bugs in removing entity aliases and determining connections for aliases
- Very old document dates now will be recognized in .jig files, but not in other types of files. We do this to avoid too many false positives.

Release 0.51 (February 2013)

- Fixed a few bugs throughout the system
- Made similarity computation more robust

Release 0.5 (May 2012)

- Capability to customize clustering (# of clusters, method) included
- Document Cluster View slightly updated for more flexible display of clusters
- Some improvements in Comma-separated value (csv) file type importer made
- More control over ordering and coloring of documents in Document Grid View
- Tablet View now can be saved and restored in a Project

Release 0.4 (May 2011)

- Document Cluster View reimplemented and improved significantly
- Comma-separated value (csv) file type importer added (now preferred over xls importer)
- Data export capabilities added to Document, List, Cluster, and Grid Views
- Regular expression entity identification added to custom entity type lists
- Alternate (white) background available for Graph and Cluster Views
- Affiliated entities added to bottom of focus document in the Document View

- User can add positive/negative words for sentiment analysis
- File importer now looks for “Date:” and “Source:” strings in top five lines of a file now to specify that meta-data
- Document summarization is faster
- Clearing all views is improved
- Document read counts and document clusters are now saved in Workspaces
- Alias pop-up in List View can be made optional now

6. Known Issues/Bugs

- The "Remove Single Entities" and “Clean Up Entities” commands may run really slowly, especially for large document collections.
- The "Export" command (to a text file) from individual views does not work in the distribution. It generates an empty file.
- The “Identify Entities” command from the Entities menu sometimes works and sometimes seems to do nothing.

7. Help/Comments

To read more about how Jigsaw works and to see a video demo, please refer to the web page http://www.cc.gatech.edu/gvu/ii/jigsaw. The web pages there, in particular the System Views page, tells more about the views. We would recommend reading the 2008 *Information Visualization* and the 2013 *IEEE Trans. on Visualization and Computer Graphics* journal papers (available at the website above) about the system for further help and explanation of Jigsaw's purpose and how it works. The overview, example scenario, and tutorial videos on the top Jigsaw web page also should be especially useful in understanding how the system and views work (although the overview video is a bit dated now). The Tutorial Videos page on the Jigsaw website has many useful how-to videos about the system.

If you would like help using Jigsaw, please send email to stasko@cc.gatech.edu and CARSTEN.GOERG@ucdenver.edu. Also feel free to call John Stasko at (404) 894-5617 if an interactive dialog would be more helpful.

We would definitely like to hear comments and thoughts about the system. We are particularly interested to hear about the way that you are using the system and if it is beneficial to you. Please do let us know about this.

8. Coming Soon (Hopefully)

Items we would like to add in new releases:

- Undo/redo
- Wikipedia import
- BibTeX import
- PubMed import
- Capturing and reviewing investigation history
- Geo-spatial View

Appendix – Jigsaw Datafile Format

Jigsaw Datafile files (with suffix .jig) are xml files that encapsulate a set of one or more documents. Presently, for each document the file contains the document ID, its date, any other documents it references, the document's source, and the actual text contents of the document, along with any entities that have been identified in the document.
A Jigsaw Datafile contains an outermost `<documents>` tag that encloses multiple `<document>` items. Each `<document>` should contain a `<docId>` and it has an optional `<docDate>` and other reference fields. The plain text source/contents of the document should be in the `<docText>` field and the identified entity values such as `<date>`, `<time>`, `<money>`, `<place>`, `<person>`, and `<organization>` trail. Note that you can add other entity types into that section as well.

There are some rules to follow for entity types, values, and other text in Project files. Entity types cannot have spaces in them. Entity values and the report description text cannot contain the &, <, >, and % characters as they are illegal in xml contents. To put those characters into text regions, use the abbreviations:

&	&amp;
>	&gt;
<	&lt;
%	&#37;

The first line of a .jig file can specify the file type, for example, Unicode, in the manner that is typically done for xml files.

An example of a Jigsaw Datafile with one document in it is shown below. Look in the `datafiles` folder for other larger examples.

```
<documents>
        <document>
        <docID>20040216-2_30</docID>
        <docDate>Feb 18 2004</docDate>
        <docSource/>
        <docText>
In the first action of its kind this winter, 18 bison were captured outside
Yellowstone National Park on Tuesday and were being tested for brucellosis.
Those that have signs of the disease will be sent to slaughter and the rest
will be marked and set free, according to Karen Cooper, a spokeswoman for the
Montana Department of Livestock.
The bison, a mix of calves, yearlings and adults, were hazed into a pen just
before noon Tuesday near Horse Butte, west of Yellowstone. The bison were then
loaded onto trailers and trucked to another holding pen to be tested for
brucellosis.
Cooper said some of the bison had been hazed back into the park on Jan. 28,
Feb. 5 and Feb. 13. "These were some of the same animals. We could not get them
back in the park so today it was a capture operation," Cooper said.
Several agencies participated in the capture, including the Department of
Livestock, Montana Fish, Wildlife and Parks, National Park Service and the U.S.
Forest Service. Through a state and federal bison management plan, government
agents haze and sometimes capture bison that leave Yellowstone. The plan is
intended to reduce the risk that bison will transmit brucellosis to cattle in
the area.
</docText>
        <date>Feb. 13</date>
        <date>Feb. 5</date>
        <date>Jan. 28</date>
        <date>Tuesday</date>
        <date>this winter</date>
        <date>today</date>
        <time>noon</time>
        <place>Yellowstone</place>
        <place>Yellowstone National Park</place>
        <person>Karen Cooper</person>
        <organization>Department of Livestock</organization>
        <organization>Montana Department of Livestock</organization>
        <organization>National Park Service</organization>
        <organization>U.S. Forest Service</organization>
        <place>Montana</place>
        <person>Cooper</person>
    </document>
</documents>
```