



## VOSON 2.0 Quickstart and Tutorial Guide

Robert Ackland (Uberlink Corporation and The Australian National University)

Francisca Borquez (Uberlink Corporation)

Jamsheed Shorish (Uberlink Corporation)

31st May 2016

This quickstart and tutorial guide covers some of the core features of VOSON 2.0. This guide assumes that you have already obtained your VOSON user account by registering on the Uberlink website, <http://www.uberlink.com>.

You can login to VOSON here: <https://voson.uberlink.com>.

Although you are now able to access VOSON through Mobile Safari, we recommend to initially use VOSON as a desktop application and to run it using a supported web browser while following along with this guide.

Some important notes before starting:

- In this guide, where you see e.g. ***Info>User*** that means you click on the *Info* menu item and select the *User* sub-menu item.
- In the bottom left hand corner of VOSON you will see *system ready* written in green text. It is important that if this is NOT shown (i.e. VOSON is working on something) that you don't give further commands to VOSON until the *system ready* text appears again.
- Throughout this guide there are references to the number of nodes in databases - these numbers may be different to what you see when you follow this guide, since website links may have changed in the meantime.

# 1. Preliminaries

After logging into VOSON for the first time, you are presented with the following menu items (See Figure 1):



Figure 1: VOSON 2.0 menu when first started.

- Info
  - **User** – This gives information on your current VOSON subscription, your access privileges and the projects that you belong to (and therefore what data are available to you). Initially, you should have access to two projects: tutorial and a project named after your VOSON username.
- Data
  - **Show databases** – This lists all the *voson* and *voson-analysis* databases that you have access to (See Chapter 3 of the User Guide for more on the different types of databases and an explanation of the fields in the *Show Databases* window). See Figure 2.
  - **Create>voson database** – this creates a *voson* database.

10 records per page Search:

Name	Project	Author	Type	Rows	Last modified	Lock status	Freshness	Parent	Tie indicator	Node Type	Action
<a href="#">testdb</a>	tutorial	tutorial	voson	381	2015-04-17_12:13:11	unlocked	n.a.	n.a.	hyperlink		
<a href="#">testdbAN</a>	tutorial	franciscab	voson-analysis	122	2007-12-14_04:12:13	unlocked	out-of-date	testdb	hyperlink		<a href="#">[delete]</a>

Figure 2: *Show databases* window.

- Help
  - Two sub-menus for accessing documentation and information about the software.

## 1.1 Tasks

- Select **Info>User** to see what your privileges are, and what projects you have access to.
- Select **Data>Show databases** to see the databases that are available to you. Initially, there should only be two databases available: *testdb* and *testdbAN*.
- In the **Show databases** window, click on the hyperlink for the *testdb* database. This opens up the tutorial database. You will see in the upper right corner of the menu bar “*testdb: 379 nodes*” (as mentioned, the number of nodes may be different). See Figure 3.

## 2. Working with voson databases: basics

You have now opened up the *testdb voson database*. This database, a crawl of 6 ANU (the Australian National University, where VOSON was created) websites, is available to all VOSON users. A *voson database* contains raw data; in this case, it contains the 379 web pages that were collected during the crawl.

You cannot conduct any analysis with a *voson database*, but you can look at and change the data. This is because VOSON separates the data (in the *voson database*) from the analysis (the *voson-analysis database*), even though both use the terminology ‘database’.

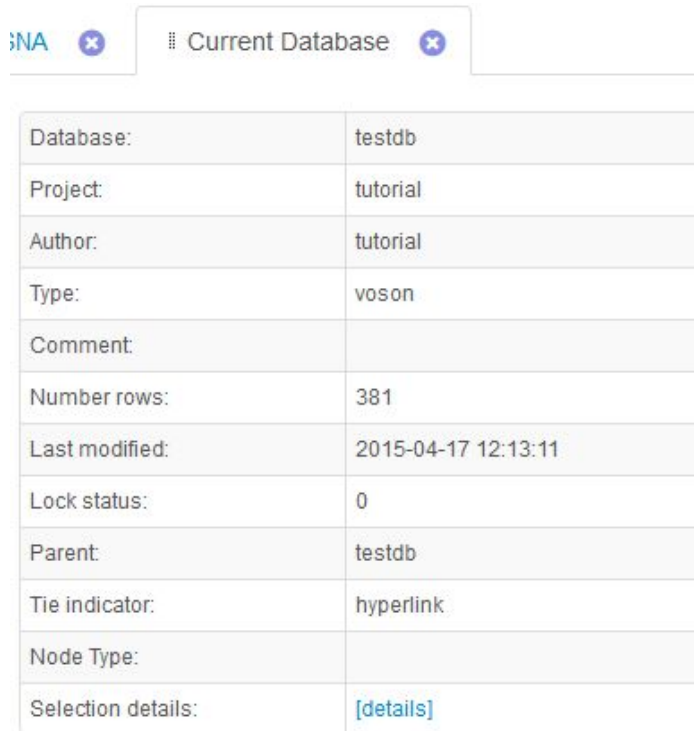
The menus that are available to you will have changed to the following (Figure 3):



Figure 3: VOSON 2.0 menu after selecting a *voson database*. Text in green on the right hand side indicates the number of nodes in the active database.

- Info

- **Current database** - this gives information on the database that is currently opened, including the project name, database name, the date of creation, the author, etc. (See Figure 4).
- **User** - as before (see the beginning of Section 1 above).



The screenshot shows a window titled 'Current Database' with a table of properties. The table has two columns: a label and a value. The properties listed are Database, Project, Author, Type, Comment, Number rows, Last modified, Lock status, Parent, Tie indicator, Node Type, and Selection details.

Database:	testdb
Project:	tutorial
Author:	tutorial
Type:	voson
Comment:	
Number rows:	381
Last modified:	2015-04-17 12:13:11
Lock status:	0
Parent:	testdb
Tie indicator:	hyperlink
Node Type:	
Selection details:	<a href="#">[details]</a>

Figure 4: Current database window.

- Data

- **Show databases** (as before).
- **DataBrowser** – this allows you to see the data, where each row is a web page. Columns indicate the identity of the page, the page URL, and other characteristics (see Figure 5).

10 records per page Search:

Row ▲	ID*	URL*	Pagegroup+	PagegroupID*	ccTLD code*	genericTLD code*	Crawl Status*	Ringset*	Pagegroup Description+
1	1	<a href="http://demography.anu.edu...">http://demography.anu.edu...</a>	<a href="http://demograohv.anu.edu.au/">http://demograohv.anu.edu.au/</a>	1	Australia	edu	1	1	0
2	2	<a href="http://econrsss.anu.edu.a...">http://econrsss.anu.edu.a...</a>	<a href="http://econrsss.anu.edu.au/">http://econrsss.anu.edu.au/</a>	2	Australia	edu	1	1	0
3	3	<a href="http://philrsss.anu.edu.a...">http://philrsss.anu.edu.a...</a>	<a href="http://ohilirsss.anu.edu.au/">http://ohilirsss.anu.edu.au/</a>	3	Australia	edu	1	1	0
4	4	<a href="http://polsc.anu.edu.au/">http://polsc.anu.edu.au/</a>	<a href="http://oolsc.anu.edu.au/">http://oolsc.anu.edu.au/</a>	4	Australia	edu	1	1	0
5	5	<a href="http://acsr.anu.edu.au/">http://acsr.anu.edu.au/</a>	<a href="http://acsr.anu.edu.au/">http://acsr.anu.edu.au/</a>	5	Australia	edu	1	1	0
6	6	<a href="http://econrsss.anu.edu.a...">http://econrsss.anu.edu.a...</a>	<a href="http://econrsss.anu.edu.au/">http://econrsss.anu.edu.au/</a>	2	Australia	edu	1	1	0

Showing 1 to 10 of 379 entries ← Previous 1 2 3 4 5 Next →

Inlinks provided by

Figure 5: Data browser.

- **Save database** – use this to save a copy of the database (note that this is different from exporting your analysis). See Figure 6.

Warning! This will write over a -voson-analysis- database with same name.

Name:  Save database

Figure 6: Save database window.

- **Add seed sites** – use this to add more seed sites to the database. Seed sites are crawled to find out who they link out to; in addition, VOSON also finds who links *into* the seed sites. See Figure 7.

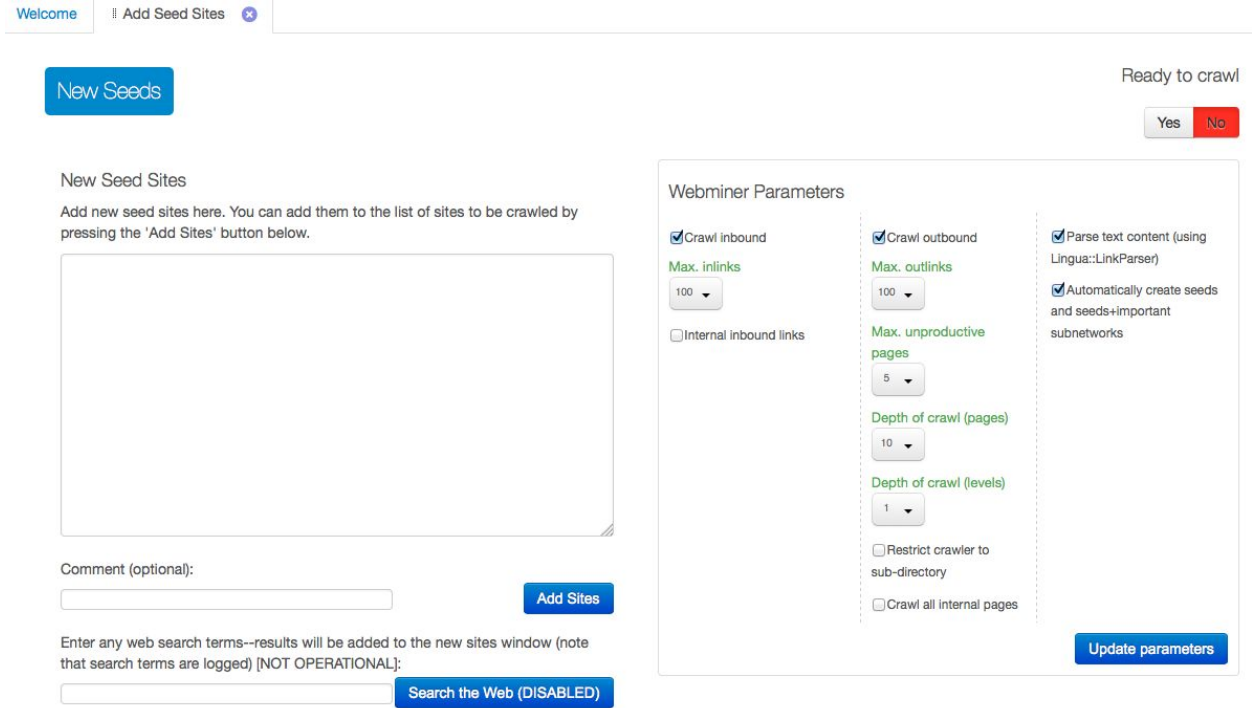


Figure 7: Add seed sites window. Webminer parameters are explained in the User Guide, section 4.5.

- **Download** – use this to export the data for viewing in other software e.g. Excel. A *voson database* can be downloaded as a comma-separated, or CSV, file (both plain text and compressed ZIP file format are supported).
- **Create**
  - **voson-analysis database** – this creates a *voson-analysis database* for the *voson database* that you currently have opened. As the name suggests, you use the *voson-analysis database* for network analysis. See Figure 8 and below for more detail.

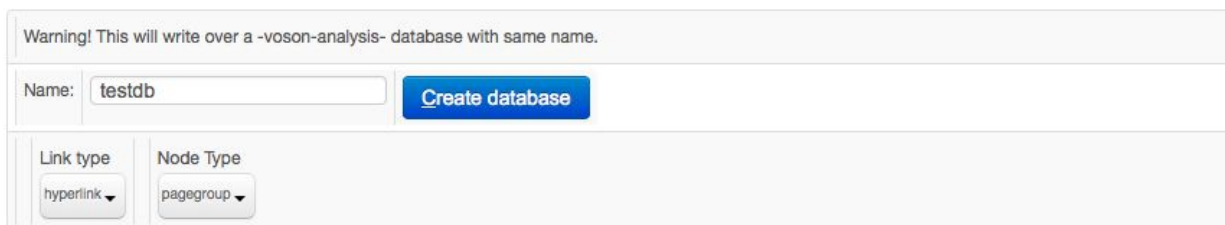


Figure 8: create a *voson analysis database* window.

- **voson database** - this creates a new *voson database*.
- **Preferences**
  - **Node pre-process>Database (Node)** – use this to influence how VOSON aggregates web pages into “pagegroups”. See section 5.1 for more details.
  - **Text pre-process>Database (Text)** – use this to pre-process the text that is collected from the meta tags in the web pages, and the text body of the web pages.
  - **Coding** – use this to create or edit categorical variables (attributes for the nodes in the network). See section 5.3 for more details.
- **Help** - As before, see the beginning of Section 1.

## 2.1 Tasks

Select **Info>Current database** to see information on the *testdb* database that you have opened.

- Select **Data>DataBrowser**. Use the navigation buttons to see what pages have been stored in the database. Scroll across to see the various information collected on each web page.
- Select **Data>Create>voson-analysis database**. A window will popup where you will be able to name a new *voson-analysis database* to create, and set the type of nodes and ties.
- Select **Data>Show databases** – your new *voson-analysis database* will now appear in your list of available databases. Open the new *voson-analysis database* by clicking on its name.

## 3. Working with *voson-analysis* databases

After opening your new *voson-analysis database*, you will notice that in the upper-right corner it says “testdb: 122 nodes”. See Figure 9 (again, the number of nodes in your database may be different to this).

Your original *voson database* contained 379 web pages (nodes), but in the *voson-analysis database*, there are only 122 nodes. This is because some pages have been merged into groups of pages, or “pagegroups”. Pages collected from the same hostname (e.g. ‘adsri.anu.edu.au’) are automatically merged together. This means that in the network map you do not have lots of network nodes simply representing individual pages from a given website. Rather, there will be a single node for that website.

With your *voson-analysis database* open, the menus will have changed to the following (see Figure 9):



Figure 9: VOSON main menu for the *voson-analysis database*.

- **Info** - as before, see the beginning of Section 1.
- **Data** - most of the available options are exactly as with the *voson database*, with the exception of the **Download** menu option.
  - **Show databases**
  - **DataBrowser**
  - **Save database**
  - **Download** – in addition to providing the CSV file format for export, four new menu items are now available for downloading *voson-analysis databases* into the GraphML and Pajek format data files.
  - **Create>voson database**
- **Analysis**
  - **Crosstab - crosstabulation**
    - **Composition** - provides crosstabulations at the node level (see figure 10).
    - **Text** - provides crosstabulations showing frequency of text (see figure 11).



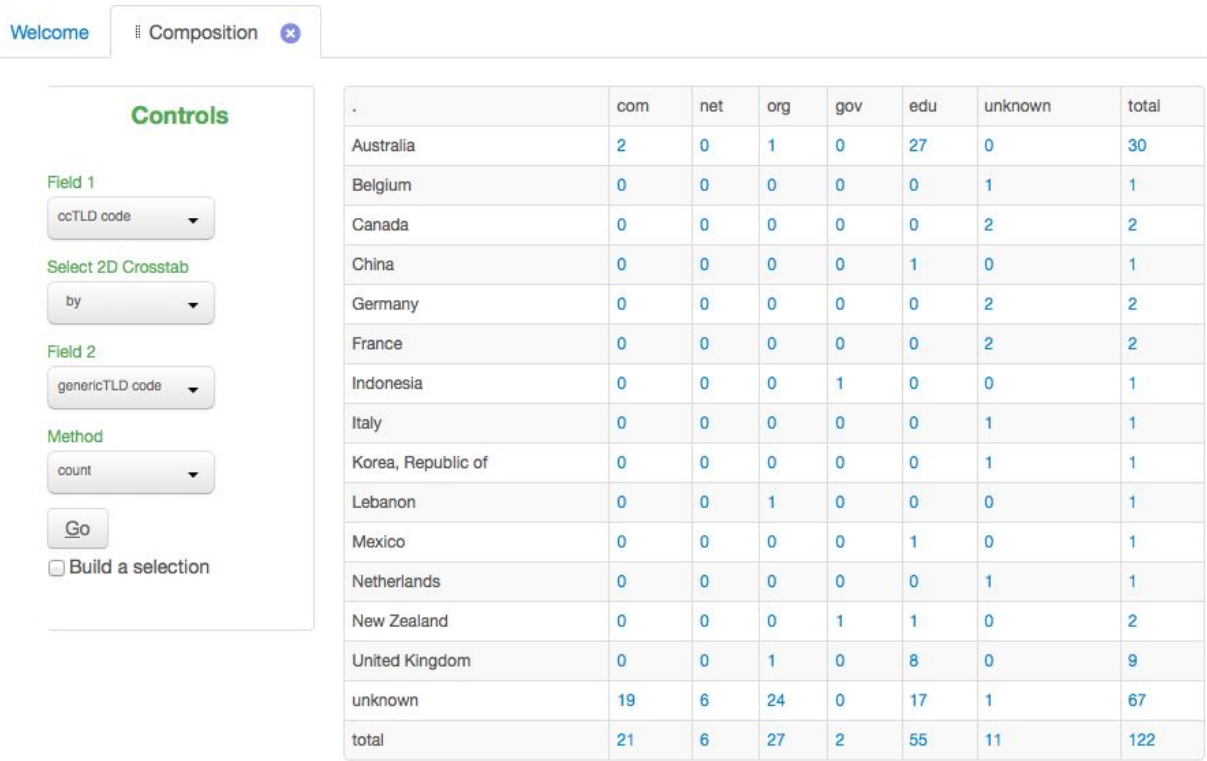


Figure 10: Example crosstabulation between the top-level domain country code, or “ccTLD code” by the generic top-level domain code, or “generic TLD code”.

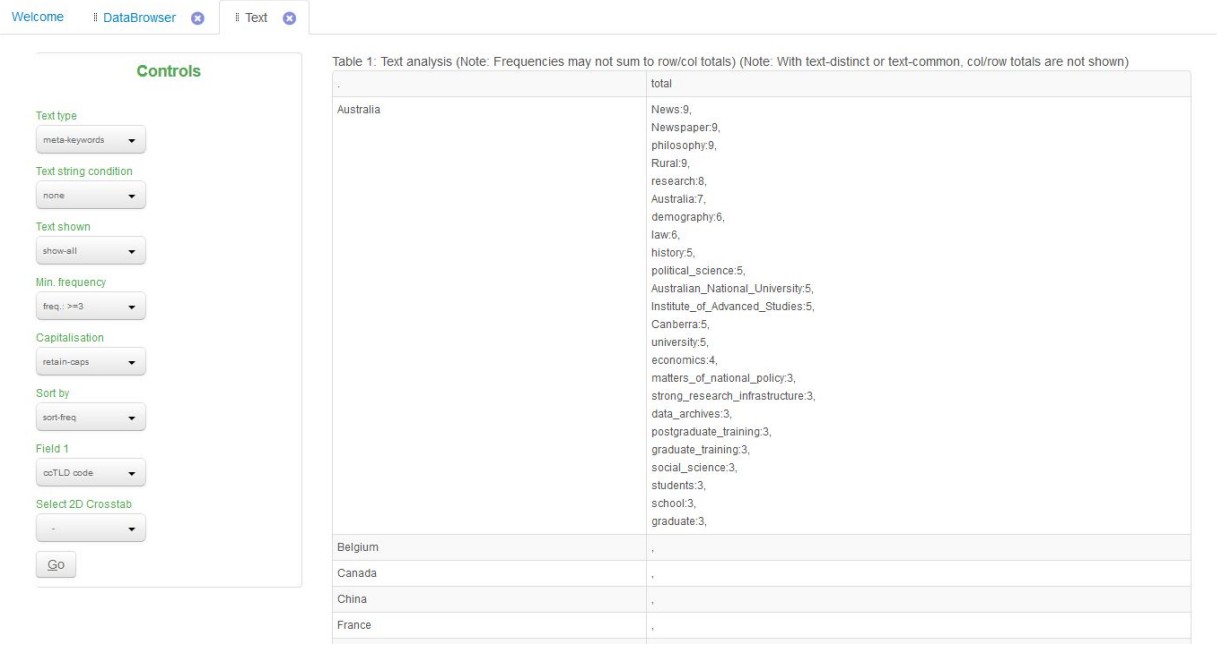


Figure 11: Example crosstabulation for text data. Crosstabulation indicates frequencies of meta keywords by Top level domain country code (ccTLD).

- **SNA** – basic social network analysis (SNA) metrics, at the level of the network (see figure 12).



The screenshot shows a window titled 'SNA' with a close button. Below the title bar is a table with 17 rows and 2 columns. The first column lists various network metrics, and the second column shows their corresponding values.

Network size	122
Number of edges	145
Number of components	1
Number of isolates	0
Smallest component size	122
Largest component size	122
Average component size	122
Number of connected nodes	122
Inclusiveness	1
Network density	0.00982252
Total number of dyads	7381
...number of mutual dyads	10
...number of asymmetric dyads	125
...number of null dyads	7246
Dyadic reciprocity 1 (ratio mutuals to all)	0.00135483
Dyadic reciprocity 2 (ratio mutuals to nonnull)	0.0740741

Figure 12: SNA metrics window.

- **Maps**
  - **Minimum spanning tree** – this is similar to an ‘ego’ network: it shows the minimum paths to/from a given root node.
  - **Complete network** – all nodes and all links are shown simultaneously (see figure 13).

- **Concept** – multidimensional scaling of text extracted from the web pages: clusterings of words possibly indicate “concepts” since they are co-occurring on web pages.
- **Hierarchy** – the nodes arranged hierarchically according to indegree, outdegree etc.

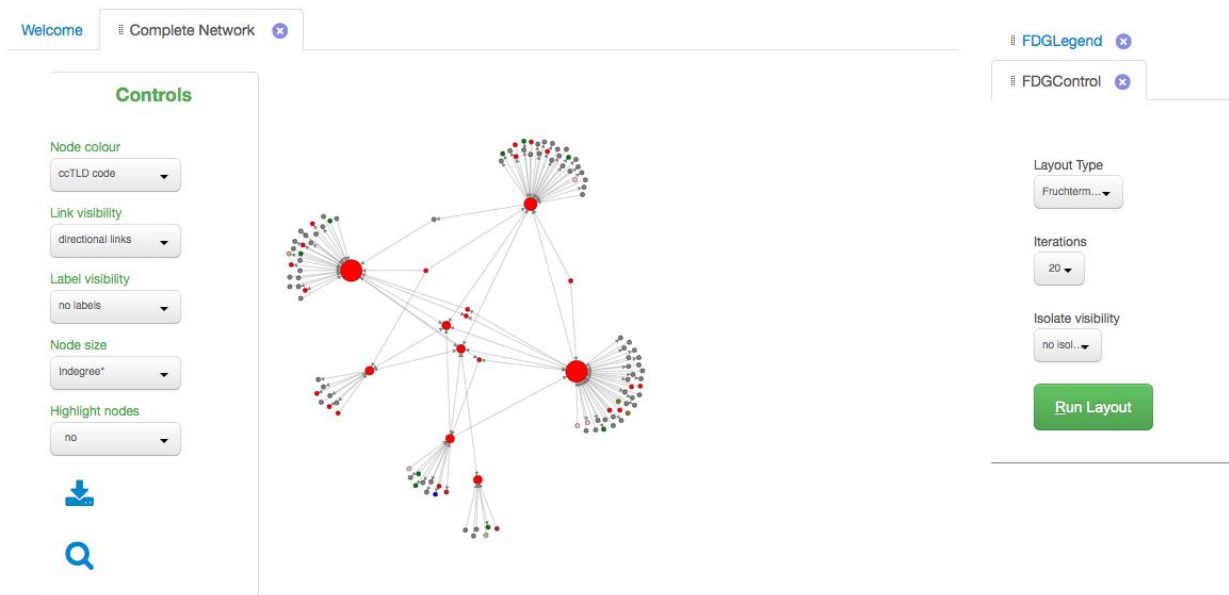


Figure 13: Example of a **Complete network** visualisation. Node colour represents ccTLD (country code). The network map displays directional links and node size by indegree. Layout type: F-R algorithm.

Maps in VOSON are interactive. Along with the controls defining node colour, link visibility, label visibility, node size and the “highlight node” function, you can zoom, pan, hover on a node to get its URL and immediate connections. In addition, you can obtain further information on a node by clicking on it and opening the *node viewer* window (figure 14).

The *node viewer* window displays all the information of the selected node, extracted from the databrowser, such as id, URL, ringset, categories, etc.

NodeViewer	
id	1
url	<a href="http://demography.anu.edu...">http://demography.anu.edu...</a>
url_pagegroup	<input type="text" value="http://demographv.anu.edu.au"/>
id_pagegrouphead	1
ccTLDc	Australia
genericTLDc	edu
crawl_status	1
ringset	1
cat1	<input type="text" value="pro"/>
cat2	<input type="text" value="Unknown"/>
cat3	<input type="text" value="Unknown"/>
cat4	<input type="text" value="Unknown"/>

Figure 14: The **NodeViewer** tab.

- **Help** – as before (see the beginning of Section 1).

### 3.1 Crosstabs - Composition

- Select **Crosstabs>Composition** – you will see the breakdown of nodes according to country code top-level domain (TLD) e.g. .au, .uk, etc. The numbers in each cell are (unscaled) frequencies, i.e. number of websites. Change the “Method” select box to “average of” and then press “Go”. The cells now show the average indegree (number of inbound hyperlinks), according to country code TLD.
- Change the “Method” select box back to “count” and then press the Go button. You will now be back at the frequency counts. Then, select the cell for Australia (30) and two

things will happen: 1) the crosstabulation will change so that it only shows Australia, and 2) in the top-right hand corner it will show “testdb: 30 nodes”.

You have just used one of the most useful functions in VOSON: you have created a subnetwork which only contains websites from Australia (.au sites).

- Now, select **Save>Database** and save this subnetwork into its own database (call it something like “testdbAust”). Look in the **Show databases** window to confirm that the new database has been saved (if you have closed it, you will need to refresh this window by selecting **Data>Show databases**).
- Open up the original *testdb voson-analysis database* again and select **Crosstabs>Composition**. Now, how do you make a subnetwork of websites from both Australia and the UK? Check the “build selection” box and select the cell for Australia – the cell will change colour. Then, select the cell for UK and it will also change colour. Do not make these selections too quickly: you need to check that you see “system ready” in the bottom right corner before making the next selection. You can now press the “Make selection” box and you will get a subnetwork containing the 38 sites from Australia and the UK.

## 3.2 Crosstabs - Text

- Select **Crosstabs>Text** – this shows a crosstabulation where the cells display the frequency of text extracted from the websites. The default is to show meta keywords, and you will see that 9 websites use the meta keyword “News”, 9 use “Newspaper” and so on. If you select “body-wordpairs” from the “Text type” select box and press the “Go” button, you will get frequencies of colocated words – these are words that are related to one another (e.g. verbs/adverbs, nouns/adjectives). The most frequent wordpair for this dataset is “financial/stress”.
- Select “meta-keywords” as the text type again (i.e. go back to the default). It is important that you do this before trying to draw a Concept map (see below).

## 3.3 SNA

- Select **Analysis>SNA** – this will give you basic SNA measures for the network. The number of nodes is 122, number of edges is 145, etc.

## 3.4 Maps – Complete network

- Select **Analysis>Maps>Complete network** – this will present a map showing all the nodes (that are non-isolates) and the connections between them, where a connection is a hyperlink.
- Change the layout algorithm from “LinLog” to “F-R” (via the *Layout Type* select box) to get a different layout.
- Change the *Node colour* select box from “ccTLD” to “genericTLD” and the nodes will become coloured according to generic TLD (e.g. .com, .edu etc.).
- Change the *Label visibility* select box to show labels (the URLs of the websites). Note that if there are many nodes then this can make the visualisation appear overly cluttered.
- Change the *Node size* select box from “none” to “indegree” – the nodes are now sized in proportion to their indegree (number of inbound hyperlinks).

## 3.5 Maps - Concept

- Select **Map>Concept** – this shows the concept maps for meta-keywords (Note: if you selected word-pairs in the **Crosstabs>Text** window, then this will not work – you will then need to go back to the Text crosstabs and change it back to “meta keywords”).
- Change the Node colour select box to “genericTLD” code – the words will then be coloured according which type of site used that word the most. For example, the meta keyword “demography” was most commonly used by .edu sites and so the word is coloured turquoise in the Concept map.

## 4. Creating a new voson database

Now it is time to create your own *voson database* containing new seed sites, and have VOSON crawl those seed sites. For the purposes of this tutorial, please use the seed sites that are listed below.

- Select **Data>Create>voson database** from the menu. Enter the new database name, e.g. “mydatabase”. Optionally, you may enter a comment describing the new database.

Note that in this tutorial, the various webminer parameters are not discussed (see the VOSON User Guide for more information), so we will use the ones that are provided by default (see Figure 15 below). Press the “Create database” button and after a few seconds you will get a popup message saying that your new *voson database* has been successfully created.

Figure 15: The *Create Database* window

- Select **Data>Show databases** and you will see your new database. You will notice that the new database currently has 0 rows. Select the new database from the *Show databases* window. Now, select **Data>Add seed sites**. This will bring up the the “Add seed sites” window (see figure 7 above). One at a time, enter the exact URLs provided below to the seed sites window and press “Add seed sites” button:

<http://vosonlab.net/>  
<http://uberlink.com/>  
<http://www.oii.ox.ac.uk/>

- Once you have entered all the seed sites, press “Yes” under "Ready to crawl" (in the upper-right hand corner). If you have enough VOSON Activity Units (VAU), then the crawl will be scheduled (see figure 16) . Note that sites might take some time to be crawled. Time will vary depending on crawl size, priority by type of subscription and on the amount of crawls on queue, among other factors. You will receive an email when the crawling has finished.

VOSON Activity Unit Estimate	
Number of seed sites	3
<b>VAU - Inbound crawl</b>	
per seed site	3.1
VAU estimate - inbound	9.4
<b>VAU - Outbound crawl</b>	
per seed site	6.4
VAU estimate - outbound	19.1
<b>Total VAU estimate for this crawl</b>	<b>28.5</b>
<b>Your current VAU allocation</b>	<b>5000</b>
<b>VAU spent so far</b>	<b>32.2</b>
<b>Estimated credit/deficit after this crawl</b>	<b>4939.3</b>

Figure 16: The VOSON Activity Units (VAU) estimate window. VAU allocation varies across subscription tiers (this is a Premium account's VAU allocation).

## 5. Data preparation

After you have received the email from VOSON saying that your data set is ready, select **Data>Show databases**, and you will see that your database now has 1363 rows. Note that this number might be slightly different for you because it depends on what the crawler found at the time it was run.

To conduct analysis, e.g. creating a network map or a crosstabulation, you need to open the *voson-analysis database*. VOSON automatically creates a default *voson-analysis database*, which uses pagegroups as units of analysis. If you wish to further preprocess your data, utilise a subset of the database or modify units of analysis, you may create another *voson-analysis database* the same way you did for the tutorial database above.

In our example, you will see that this database shows the online network formed by the VOSON project's site (Australian National University), Uberlink's site and the Oxford Internet Institute's site.

Next, we are going to look at three of the data preparation features in VOSON. Note: At present, data preparation can only be done when you have a *voson database* open. So, while you might use the *voson-analysis database* to identify the data preparation that you need to do (for example, by looking at the network maps), you need to open up the corresponding *voson database* to do the actual data preparation.

Before starting with the data preparation, it is best to first create a copy of the *voson database* using **Data>Save database** (call it something like "mydatabase2"). You need to then create a *voson-analysis database* corresponding to this copy of the *voson database*.



## 5.1 Node pre-processing

For the three node processing tools (*pagegrouping*, *pruning* and *preserving*), we recommend to use the syntax provided in these examples.

### 5.1.1 Pagegrouping

If you open up the *voson-analysis database* and look at the databrowser, you will notice that there are sites that are subsites of the OII's main site, e.g.:

<http://elections.oii.ox.ac.uk>, <http://cii.oii.ox.ac.uk> and <http://rural.oii.ox.ac.uk>.

The OII's web presence would be more accurately represented if these sites were merged together with the main site. This process is called “pagegrouping” - it is altering the way VOSON creates pagegroups (aggregations of web pages).

To do so, open up the *voson database* and select **Preferences>Node pre-process>Database** (See Figure 17) and in the *pagegrouping* box enter the following:

**`www.oii.ox.ac.uk,elections.oii.ox.ac.uk,cii.oii.ox.ac.uk,rural.oii.ox.ac.uk`**

The screenshot displays three distinct sections for node pre-processing:

- Pagegrouping:** Features a text input field labeled "Add new pagegroup set:" with an "OK" button. Below it, a row shows "D: DummyWordIgnore,DummyWordIgnore" followed by a small "R" button.
- Pruning:** Includes a text input field labeled "Add new pagegroup:" with an "OK" button. To the right, there is a section titled "Prune non-navigable content" with "Yes" and "No" buttons. Below this, a row shows "D: DummyWordIgnore" followed by a small "R" button.
- Preserving:** Features a text input field labeled "Add new pagegroup:" with an "OK" button. Below it, a row shows "D: DummyWordIgnore" followed by a small "R" button.

Figure 17: Node preprocessing tools, including pagegrouping, preserving and pruning commands.

Then re-create the *voson-analysis database*. You will see that it now has 492 rows compared with 496. When you create the complete map you will see that <http://elections.oii.ox.ac.uk>, <http://cii.oii.ox.ac.uk> and <http://rural.oii.ox.ac.uk> are no longer in the map – they have been merged into <http://www.oii.ox.ac.uk>.

### 5.1.2 Pruning

Quite often you get websites being picked up by the crawler that are irrelevant for your study. A good example is <http://www.adobe.com> – this gets picked up because sites link to the Acrobat Reader.

To remove these sites use the “pruning” function, from **Preferences>Node pre-process**, when you have a *voson database* open. To prune <http://www.adobe.com> from this network (if it were present), we would enter “**adobe.com**” into the “pruning” box, and then re-create the *voson-analysis database*.

### 5.1.3 Preserving

As mentioned before, the default VOSON seed processing algorithm aggregates pages from a given host into a single pagegroup (or site). For that reason, you may get unrelated pages or individual sites representing different organisations that share the same domain being grouped together into a single node.

For example, you may have two or more Wikipedia or Twitter entries being aggregated into a single node, when in fact it might be better if they are kept separate.

If you open the *voson database* and then the *Data Browser*, you can search for “Twitter”, and you will see that there are some pages from Twitter in the database. The majority of the URLs are Tweets of the OII, therefore it would make sense to keep them grouped together. One of them, however, is a Tweet by a user who is mentioning OII. If we consider that URL as a node representing a social actor, then we would like to make that URL independent from the pagegroup or “preserve” it. The URL in this example is:

<https://twitter.com/search?q=%23OxDeg>

With the *voson database* open, select **Preferences>Node pre-process** and in the *preserving* box enter the following (without the ‘https://’ prefix):

[twitter.com/search?q=%23OxDeg](https://twitter.com/search?q=%23OxDeg)

After entering the above, create the *voson-analysis database* again and compare the number of rows with the previous *voson-analysis database*. The new database has one more row.

## 5.2 Text pre-processing

In order to get useful text analysis, you generally need to do some text pre-processing. Although this is not covered in this tutorial, you may find comprehensive information in the VOSON User Guide.

## 5.3 Coding

VOSON automatically creates two categorical variables: country code TLD and generic TLD, which you've seen while following this tutorial. While useful, researchers will often need to create their own categorical variables so the network maps can have node colour schemes that are appropriate for their research.

In this example, we will create a new categorical variable that will be used to indicate whether or not a website is focused on Internet research. With the *voson database* open, click on **Preferences>Coding**. Press the left add "+" button and a new categorical variable "cat1" will appear. Press the pencil icon to edit this new variable, and the details will appear on the edit panel on the right hand side. The default variable label is "new categorical variable – 1" - change this to "Internet focus" and press the "Save" button. Initially, there is only a single value 0, for unknown. Press the "Add" (+) button in the edit panel "Categorical value successfully added". See figure 18.

Note: the new categorical value will not be shown in the edit panel – you need to press the pencil icon for it to be shown.

The new value will have a default value label of "label for value 1". Change this to "focused", and change the colour to red. Press "Save". Then press the "Add" button again (followed by the pencil icon) and change the label for value 2 to "not focused" and the colour to blue. Press "Save" again.

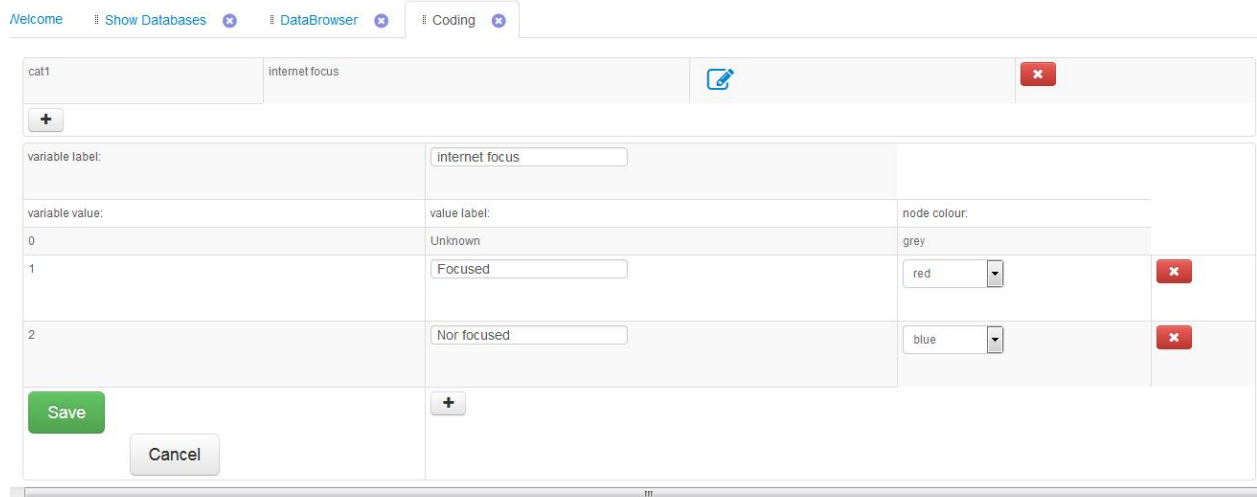


Figure 18: The Coding window.

Congratulations: you have just created a categorical variable! You'll need to re-open the *voson database* before you can see it. After re-opening the database, if you go to the DataBrowser, you will see the new categorical variable as a select box. You can code the sites from the DataBrowser by choosing appropriate values of the select box for each site. (You use the DataBrowser to code the sites.)