Importing Data into ORA

Jeff Reminga
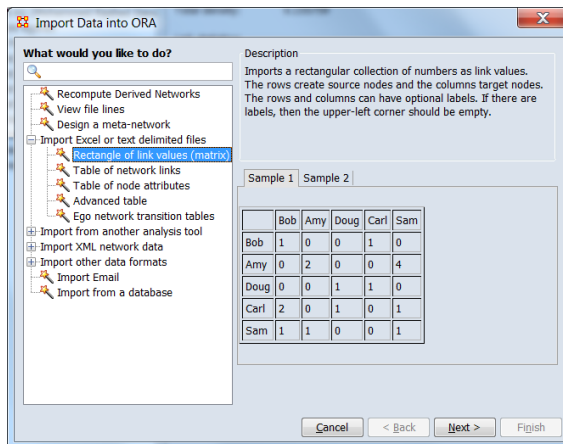
The CASOS Center
COS Program, School of Computer Science, Carnegie Mellon
Summer Institute 2020

Carnegie Mellon

**Center for Computational Analysis of
Social and Organizational Systems**
http://www.casos.cs.cmu.edu/

Carnegie Mellon

Import Wizard: Matrix data

A common data format for network data is a matrix: an arrangement of data into cells with rows and columns.

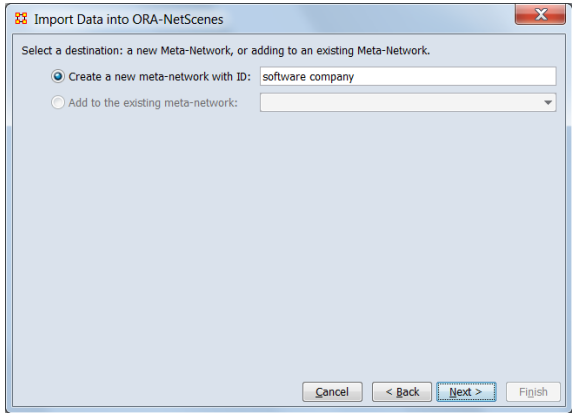Each cell (i,j) stores a numerical value associating row i with column j.

The matrix can optionally have row and column headers.

Open the data import wizard. Select the option shown.

<Your Name>

Import Wizard: Matrix data...

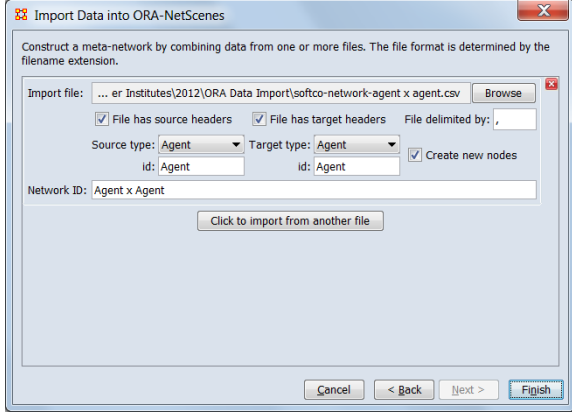Click Next and then specify an ID for the meta-network.

I chose to import into an new meta-network called software company.

As I have data in CSV (comma separated value) files for this company that we will practice importing.

Import Wizard: Matrix data...

Click Next and then Browse to the specified file.

I know the file contains Agent x Agent data, and so the source and target types are changed.

We have row and column headers (labels).

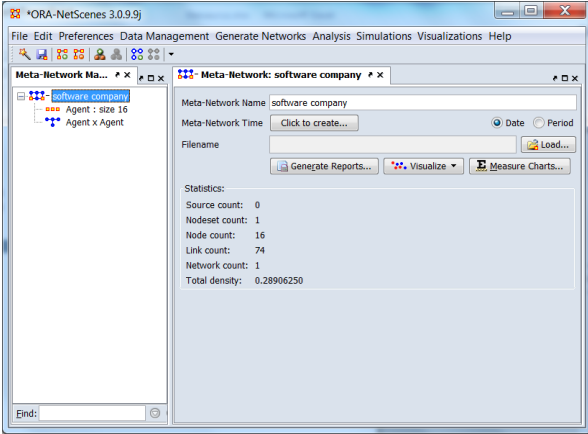Our file is comma (,) delimited.

And we want to create new nodes for each label.

<Your Name>

Slide 5

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Import Wizard: Matrix data…

The meta-network manager should have the entry shown.

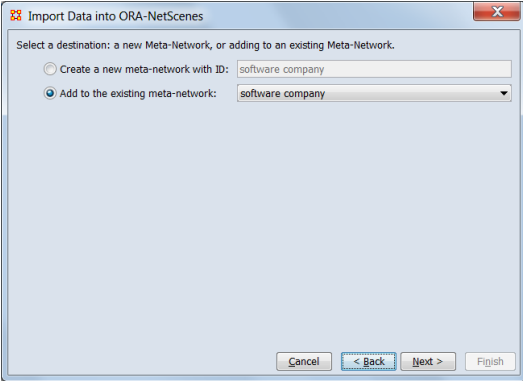You can select the agent nodeset and examine the nodes in the editor.

And likewise for the network Agent x Agent.

June 2020   © 2020 CASOS, Director Kathleen M. Carley   5

Slide 6

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Import Wizard: Matrix data…

We have more networks associated with this software company. They are in separate files.

We can import them into our existing software company meta-network as follows:

1. Open the Data Import Wizard
2. Select the same import method as before
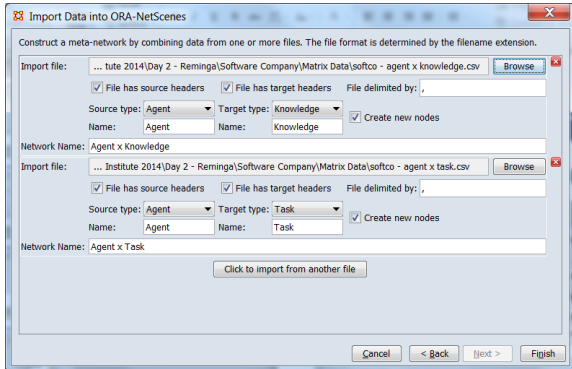3. Select to add to the existing meta-network (as shown)

June 2020   © 2020 CASOS, Director Kathleen M. Carley   6

<Your Name>

Import Wizard: Matrix data...

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

We will import two files at once:

Agent x Knowledge
Agent x Task

Use the "Click to import from another file" button.

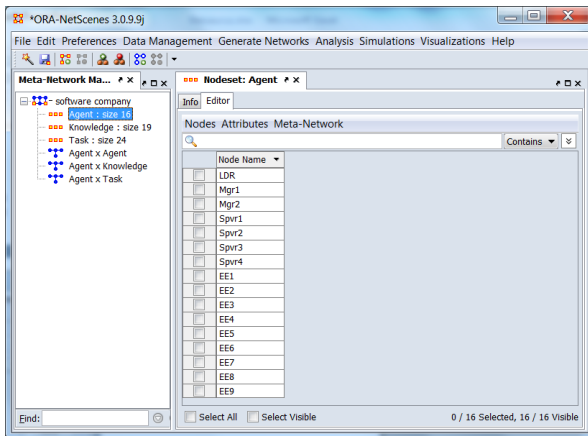Be sure to choose the Source and Target types for each file to import.

Then click finish.

Import Wizard: Matrix data...

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

The meta-network manager should now have these nodesets and networks.

Notice that we have only the node name, and no Agent attributes.

Perhaps we have more information about the agents, such as Gender, Experience, Age, etc.
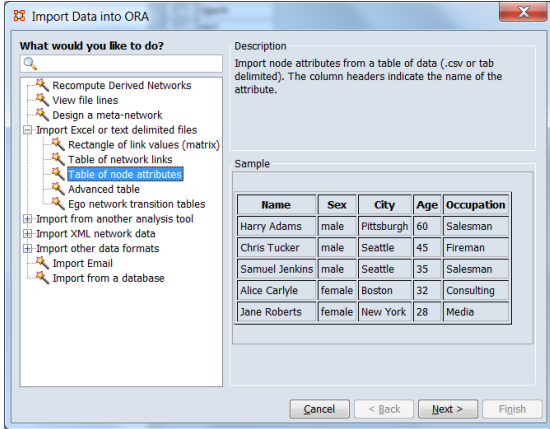
And we will next demonstrate how to import such data.

Import Wizard: Node Attributes

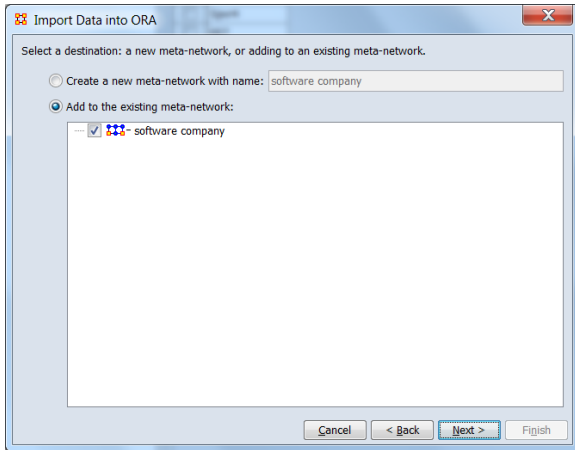Open the Data Import Wizard and choose the option shown.

A Node Attribute is a characteristic or feature associated with a node.

For example: agent nodes could have a text attribute called "hometown", a numerical attribute called "age", and a categorical numerical attribute called "experience level". This will be further a little later.

June 2020   © 2020 CASOS, Director Kathleen M. Carley   9

Import Wizard: Node Attributes...
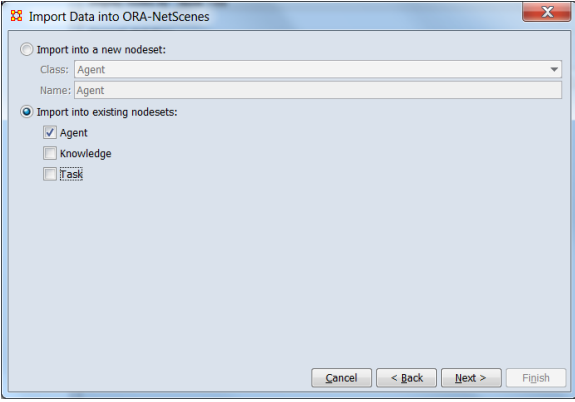
Click to add the attributes to our software company, as shown.

June 2020   © 2020 CASOS, Director Kathleen M. Carley   10

Import Wizard: Node Attributes…

We will be adding attributes to an existing nodeset.
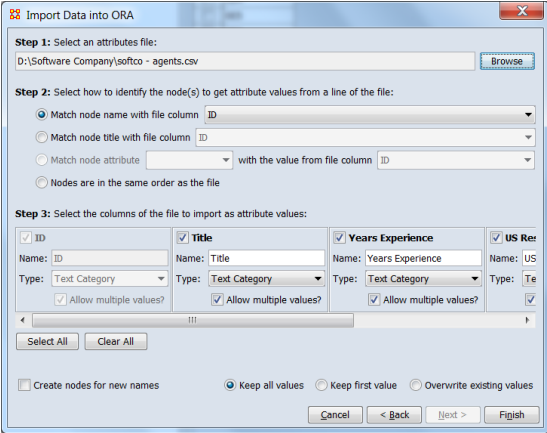
Namely, we will add them to the Agent nodeset.

June 2020 — © 2020 CASOS, Director Kathleen M. Carley — 11

Import Wizard: Node Attributes…

Click Next and then Browse to the specified file.

Step 2: indicate how to match up the attribute data to the existing nodes. We have a Name column to identify the agent node, and the row contains attributes for that agent.

Step 3: we will use all of the attributes, and we need to tell which data type to use.

Indicate whether to overwrite or accumulate values.

June 2020 — © 2020 CASOS, Director Kathleen M. Carley — 12

<Your Name>

Node Attribute Details

- A Node Attribute is data associated with a node
- An attribute has a **Name** that must be unique (case-insensitive) within the nodeset
- An attribute's **Data Type** determines the operations that ORA performs with the values:
  - Numerical distributions can be computed using Number values
  - Grouping nodes can be computed using Categorical Text or Number values
  - Whether a value is Number of Number Category is a user decision based on what the value represents and how it should be used

Node Attribute Data Types

- **Data Types** for an attribute:
  - **Text**: value is an arbitrary text string, a "blob" of text
  - **Text Category**: value is from a collection of strings, such as "Country" or "Color"
  - **URI**: value that represents a link to other data (a URI is a generalization of a URL
  - **Number Category**: value from a set of numbers, such as "Experience Level", or even "Age"
  - **Number**: value is from a distribution, and statistics are to be computed across node values, such as "Strength" or "Income"

<Your Name>

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Node Attribute Multiplicity
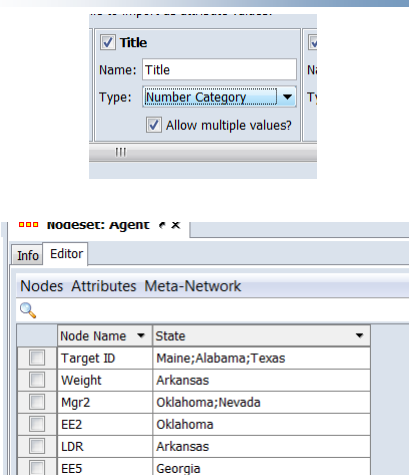
- A Node Attribute can be **Single-valued** or **Multi-valued**
- Single-valued is more memory efficient and faster
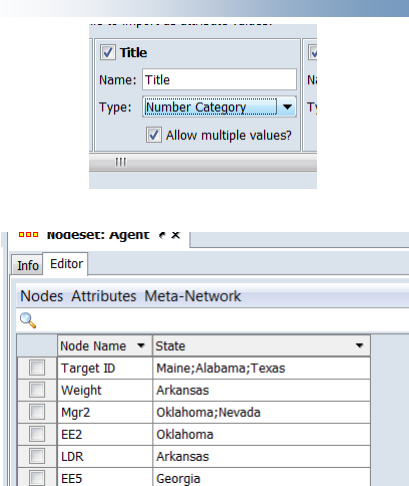- Multi-valued appears in the Nodeset Editor as a semi-colon separated list of values

CASOS

June 2020   © 2020 CASOS, Director Kathleen M. Carley   15

Carnegie Mellon
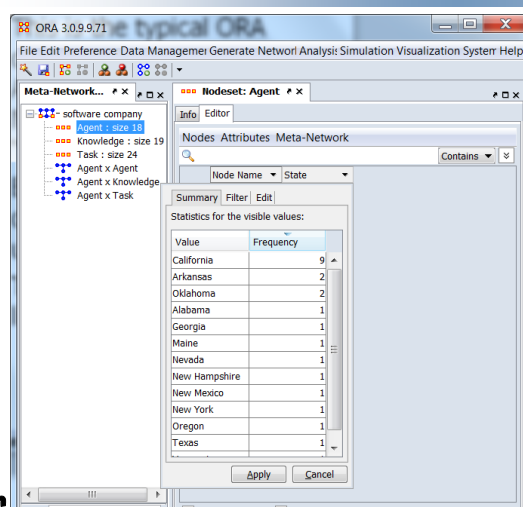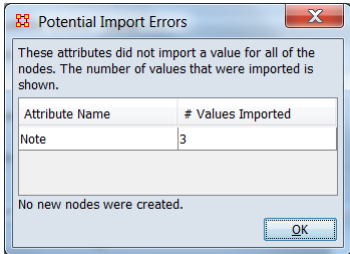isr institute for SOFTWARE RESEARCH

Node Attributes in the Interface

☑ Title
Name: Title
Type: Number Category ▾
☑ Allow multiple values?

Nodeset: Agent
Info | Editor
Nodes Attributes Meta-Network

Node Name ▾	State
Target ID	Maine;Alabama;Texas
Weight	Arkansas
Mgr2	Oklahoma;Nevada
EE2	Oklahoma
LDR	Arkansas
EE5	Georgia

- This is the typical ORA control for creating a new Attribute: notice the Name, Type, and "multiple values"
- Enter multiple values into cells in the Nodeset Editor using a semi-colon
- The tool-tip for the cell lists the values separately

CASOS

June 2020   © 2020 CASOS, Director Kathleen M. Carley   16

Slide 17

Node Attribute Controls

- This is the typical ORA control for creating a new Attribute: notice the Name, Type, and "multiple values"
- Enter multiple values into cells in the Nodeset Editor using a semi-colon
- The tool-tip for the cell lists the values separately

Slide 18

Node Attribute Controls…

- In the Nodeset Editor you can view the attributes by clicking on the drop-down
- Summary shows values – depends on the data type
- Filter lets one select nodes to show
- Edit allows one to change data type and Name

<Your Name>

Import Wizard: Node Attributes…

ORA gives feedback on potential import problems.

In our case, the attribute **Note** did not import a value for each node. Only 3 nodes received a value during import.
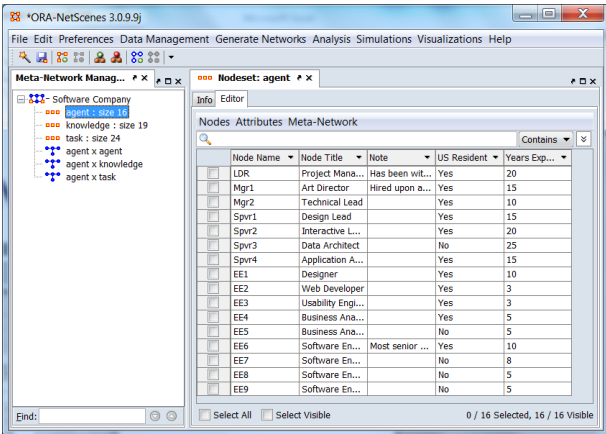
This is expected, since our node attribute file indeed only has three values to import.

19

Import Wizard: Node Attributes…

After the import, the Agent nodeset has the imported attributes.

We can now save our software company via File \ Save.

Note that this saves all of the data (nodes, networks, and attributes) into a single file in an XML format called DyNetML.

It can be loaded back into ORA with File \ Open.

20

<Your Name>

<Your Name>

Import Wizard: Tables...

Browse to the specified file in the Table Link folder.

Step 2: there are 3 columns in the data file: Source (node) ID, Target (node) ID, and (link) weight.

We specify that the columns contain Node names, both for nodeset Agent. For the Weight column we do not need to specify anything.
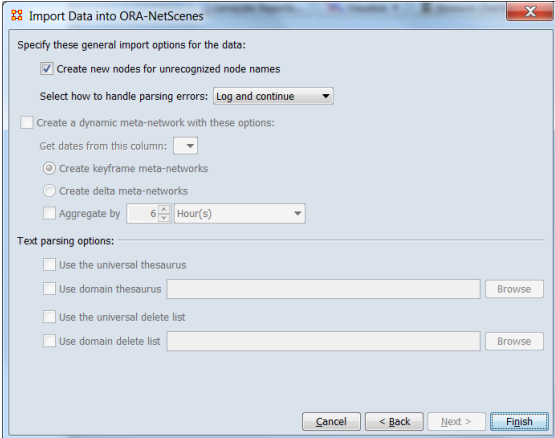
Step 3: create an agent x agent network by associating the two ID columns, and assign the link the value in the weight column.

We check to create new nodes.

Import Wizard: Tables...

- After clicking next, some advanced options appear
- We will use the default options
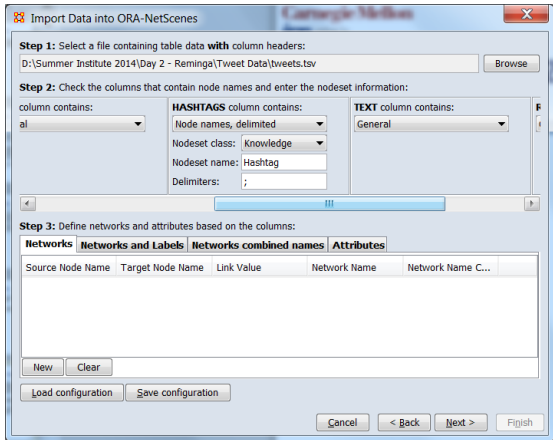- Just click Finish

<Your Name>

Import Wizard: Tables...

The resulting meta-network is created.

It should look identical to that of the previous import.

June 2020   © 2020 CASOS, Director Kathleen M. Carley

Import Wizard: Delimited Tables

As another example, we will import the Tweets.csv file from the Tweet Data folder.

The HASHTAG column contains multiple hashtag node names, delimited by a semicolon, and we indicate this in Step 2.

June 2020   © 2020 CASOS, Director Kathleen M. Carley

Import Wizard: Data over time

- Often data is associated with a point or period in time
  - A tweet occurs at a particular time
  - A longitudinal survey produces a sequence of Agent x Agent communications networks, one per month for a year
  - Email data and telephone call data are time stamped
- ORA's Dynamic Meta Network is a collection of Meta Networks, one per time period
  - Each meta-network is called a keyframe
  - Each meta-network has a unique time stamp date or string
- Measures and reports analyze Dynamic Meta Networks to find patterns and trends in the data over time

CASOS

Import Wizard: Data over time...

- Data tables to import will often have a column that contains the date or time that something occurs
- The Data Import Wizard contains support to use such a date/time column and automatically create a Dynamic Meta Network
- Each node and link is put into the Meta Network corresponding to its timestamp

CASOS

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Import Wizard: Data over time...

- The Data Import \ Tweet Data folder contains a file called tweets.tsv
- Each row records a tweet
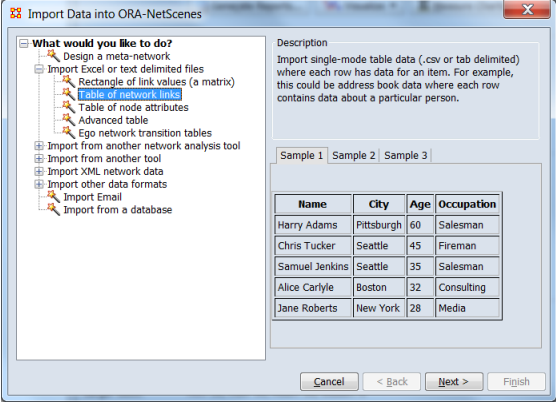- There is a DATE column that we will use to import the tweets and create a Dynamic Meta Network

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Import Wizard: Data over time...

- Open the Data Import Wizard and choose the option:

CASOS

<Your Name>

Import Wizard: Data over time…

Step 2 indicates that the column has Dates, and that they are formatted yyyy-mm-dd
Ex: 2016-01-23

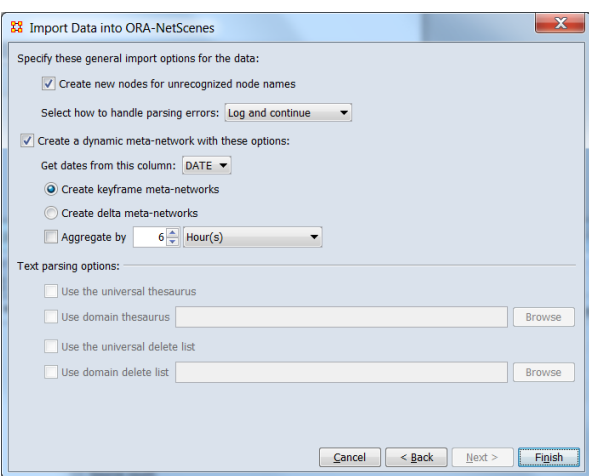This will create one meta-network per day with the nodes and links for that day.

Note: you could "truncate" the date format to yyyy-mm to create and populate one meta-network per month

Import Wizard: Data over time…

Click Next and enter information about using the Date column.

If you had multiple Date columns you could select which to use as dates.

Create keyframes if you want "time buckets" that partition your data into periods.

Create deltas to think of your data as incremental additions of data.

Use the **Aggregate by** to combine meta-network periods into days, months, or years.
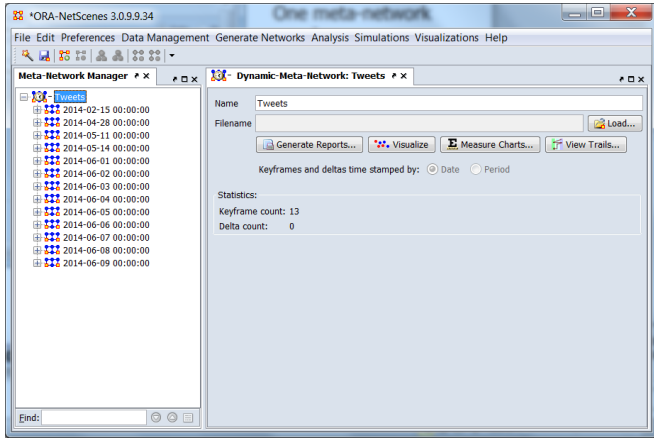
<Your Name>

Import Wizard: Text Tables
(Only available in Full Version)

When a column of data contains unstructured text, such as the message from a tweet, use Step 2 to indicate this for the column.

The importer will create a node for each word found in the text message and add the node to the nodeset with class Unknown.

The nodeset class is unknown because there is no import information for what type of node it is: Agent, Location, Knowledge, Resource, Task, etc.

A thesaurus is a file that maps a word to its nodeset class, for instance, Germany to Location, or Eisenhower to Agent.

A delete list is a file that contains a list of words to ignore during import.

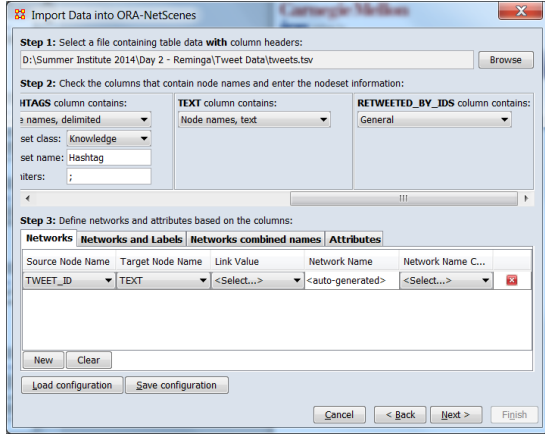We can use both a thesaurus and delete list during import.

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Import Wizard: Text Tables...

Use the Tweets.csv dataset within the Tweet Data folder.

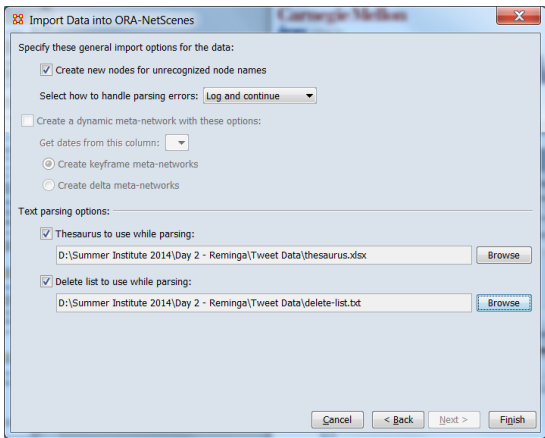The Text column contains the text message from a tweet.

Use the settings shown in Step 2.

Use Step 3 to create a network that maps the Tweet to each of the word nodes parsed from the Text column.

CASOS

June 2020   © 2020 CASOS, Director Kathleen M. Carley   35

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Import Wizard: Text Tables...

Click Next.

The advanced option page contains a "Text parsing options" section where you can choose a Thesaurus and a Delete List to use during the import.

CASOS

June 2020   © 2020 CASOS, Director Kathleen M. Carley   36

CASOS

Slide 37

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Import Wizard: Text Tables...

	A	B	C	D
1	conceptFrom	conceptTo	metaOntology	nodeType
2	world cup	world_cup_2014	Event	specific
3	world cup 2014	world_cup_2014	Event	specific
4	john oliver		agent	specific
5	england		location	generic

A thesaurus file needs to have at least these 4 columns. This is the same format that Automap uses for a thesaurus.

Using this thesaurus would create a "world_cup_2014" Event node, marked with a NodeType = Specific attribute whenever it found "world cup" in the string.

Occurrences of "John Oliver" would be made into Agent nodes "John_Oliver".

CASOS

Slide 38

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Import Wizard: JSON

JSON files can be imported by using the option shown.

CASOS
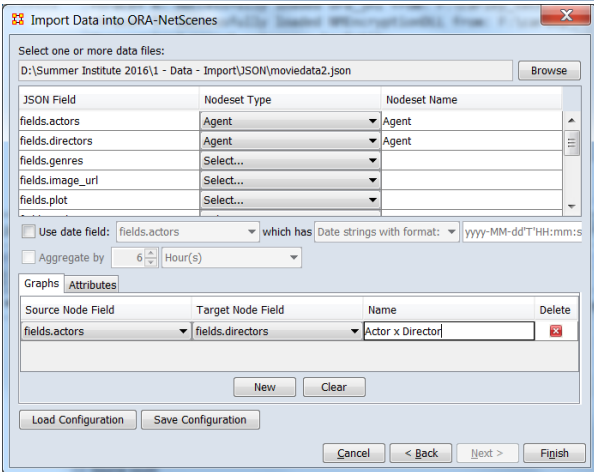
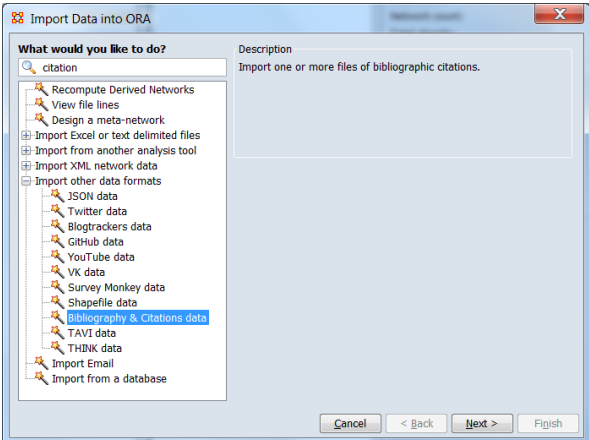**CASOS**

<Your Name>

Import Wizard: JSON…

Click Next, and then browse to the movedata2 file that is in the JSON folder.

This contains information about movies. Each movie is a JSON record.

In this example, we will associate each actor in a movie with his director.

June 2020 © 2020 CASOS, Director Kathleen M. Carley 39

Import Wizard: Bibliography & Citation data

- Bibliographic, or citation data, is created when an author cites the work of others in a bibliography
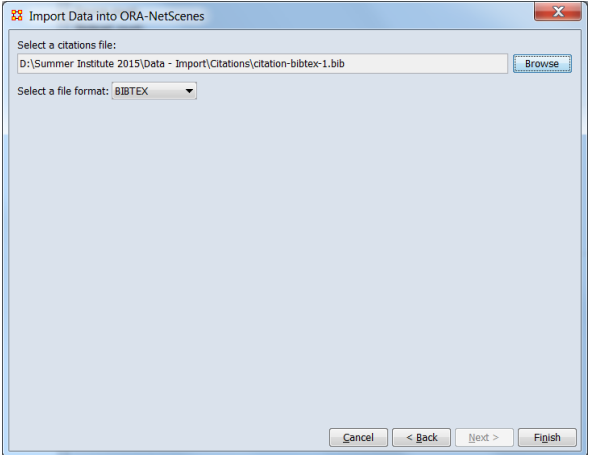- When an author cites another author, a link is created to form a citation network
- The data are usually found in these formats: Scopus, RIS, or BibTex

June 2020 © 2020 CASOS, Director Kathleen M. Carley 40

<Your Name>

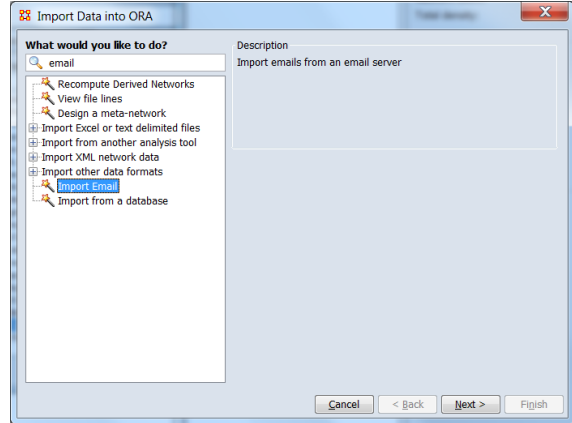**Import Wizard: Bibliography & Citation data…**

- Simply select the file that contains the citation data
- Select the correct data format: BibTex, RIS, Scopus CSV
- Click Finish
- Refer to Help for a complete description of the networks created.

June 2020 © 2020 CASOS, Director Kathleen M. Carley 41

Import Wizard: Email data

Email data be imported from an IMAP server, which is very common.
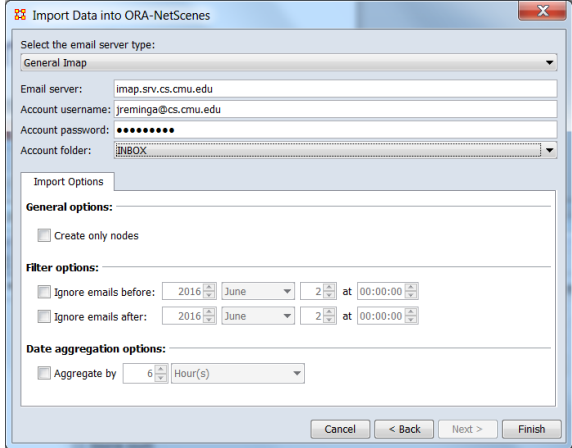
Gmail has an IMAP server.

Select the option shown.

June 2020 © 2020 CASOS, Director Kathleen M. Carley 42

<Your Name>

Import Wizard: Email data...

Click Next and enter the email server and user account information.

Then click on folder Select... to verify the connection and get possible email folders.
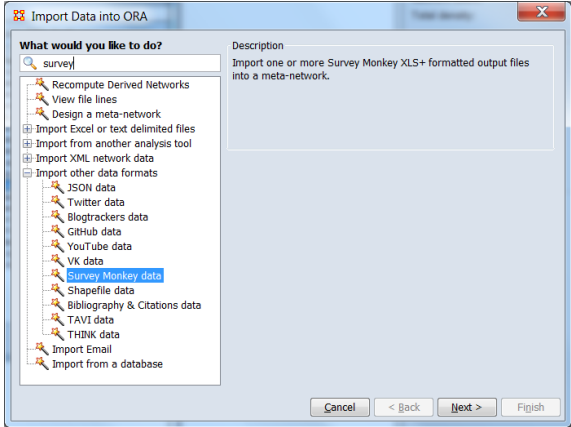
Use the Import Options tab to filter emails based by date.

Emails can be imported into time periods by using the Aggregate control.

Import Wizard: Survey Monkey

Survey Monkey files can be imported using the option shown.
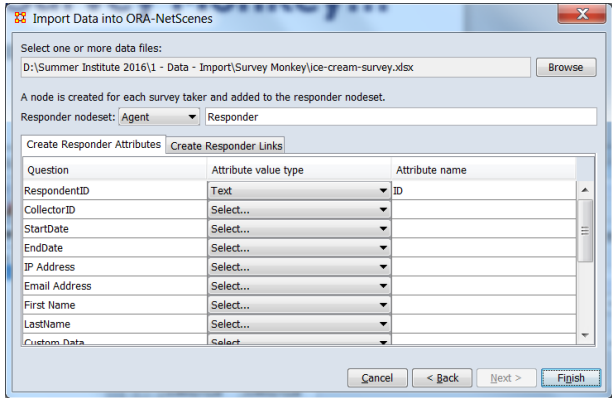
Only survey data files in the XLS+ format can be imported.

The files must contain the actual response values.

Import Wizard: Survey Monkey…
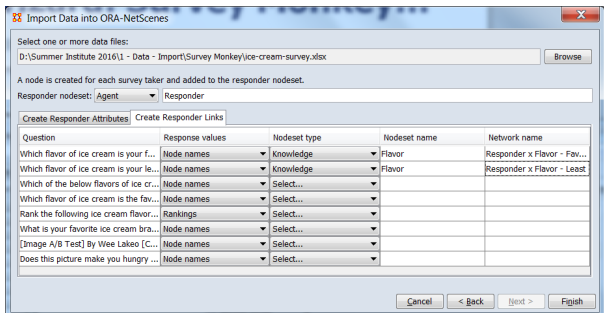
Browse to the file:
ice-cream-survey.xlsx

The survey contains one row per survey taker, called a respondent.

Each respondent is imported as a node, along with the attributes specified in the Create Responder Attributes tab.

Import Wizard: Survey Monkey…

Links for the responder are specified using the Create Responder Links tab.

For each survey question, first specify using Response Values the type of data in the response.

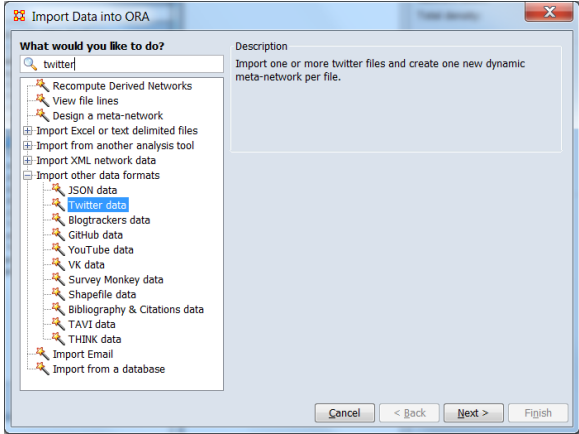Then specify the target nodeset, and a network type.

Here we create a target nodeset of Flavor, and put the links in different networks (Favorite and Least Favorite) based on the question.

Slide 47

Data Import: Twitter Data

- Every tweet is recorded by Twitter
- Twitter makes this data available in a text format called Twitter JSON
- Twitter JSON data can be imported into ORA using the Data Import Wizard

Slide 48

Data Import: Twitter Data...

- There are many import options for Twitter data
- **General options:**
  - Create only nodes and attributes without links and networks
  - Use this save space or to get a general overview of the tweets, tweeters, topics and locations
- **Filter options:**
  - Choose to only parse tweets in a particular date range
  - And to ignore tweets that contain a particular word
  - Or to only accept tweets that contain a particular word

<Your Name>

Data Import: Twitter Data...
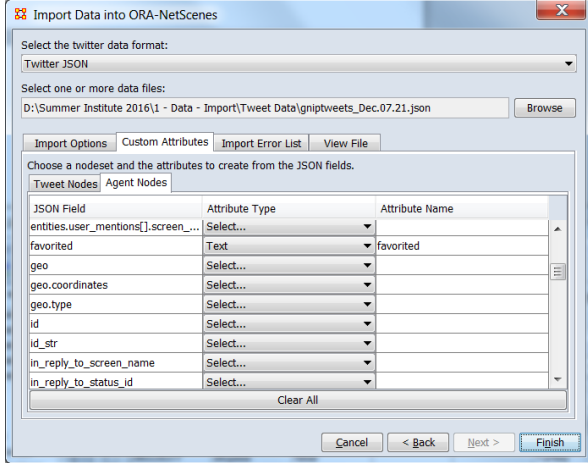
Use the Custom Attribute tab to import custom attribute values from the JSON.

Data Import: Twitter Data...

- **Tweet message options:**
  - Choose whether to store the entire tweet message as a tweet node attribute or not (to save space)
  - Parse the tweet message using text analysis techniques *(available only in the Full Version)*
  - Clean the tweet message using a thesaurus *(available only in the Full Version)*

Carnegie Mellon
institute for
SOFTWARE
RESEARCH

Data Import: Twitter Data...
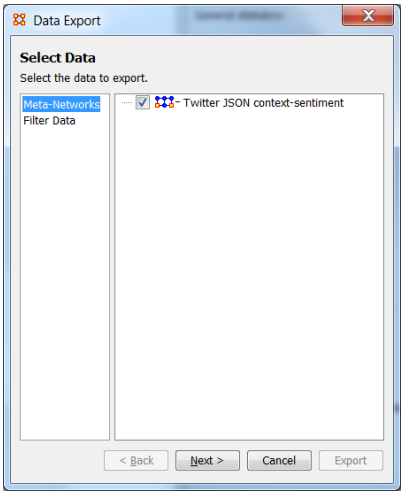
- **Date aggregation options:**
  - Choose to combine tweets together into time periods (for example: group tweets by month)
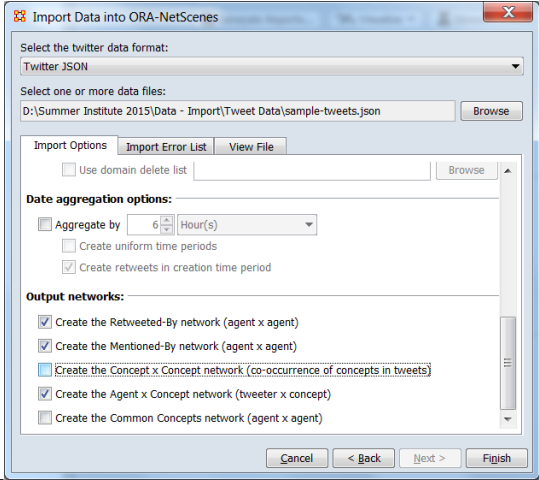- **Output networks:**
  - Select which networks to create
  - Select only the ones you want in order to save space

Carnegie Mellon
institute for
SOFTWARE
RESEARCH

Data Export

It is easy to save the meta-network to different formats, such as: CSV, UCINET (DL, binary .##h, and VNA), Pajek, and Analyst Notebook.

Use the main menu item:
File \ Data Export

Select the meta-networks to export.

Use the Filter Data page to select which nodesets and networks to export.

Click Next.

Data Export...
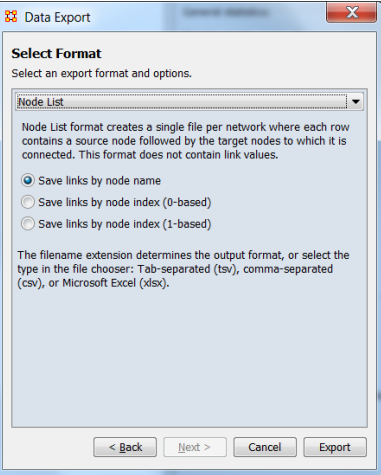
Select Format
Select an export format and options.

Node List ▾

Node List format creates a single file per network where each row contains a source node followed by the target nodes to which it is connected. This format does not contain link values.

◉ Save links by node name
○ Save links by node index (0-based)
○ Save links by node index (1-based)

The filename extension determines the output format, or select the type in the file chooser: Tab-separated (tsv), comma-separated (csv), or Microsoft Excel (xlsx).

< Back   Next >   Cancel   Export

Here the Node List option is selected which is a very compact format.

There are different options for how to save out the nodes.