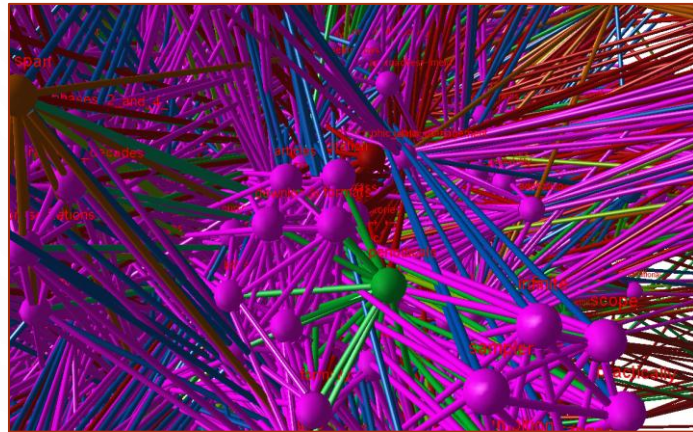


# AUTOMAP-PING TEXTS FOR HUMAN-MACHINE ANALYSIS AND SENSE-MAKING



**SHALIN HAI-JEW**

**KANSAS STATE UNIVERSITY**


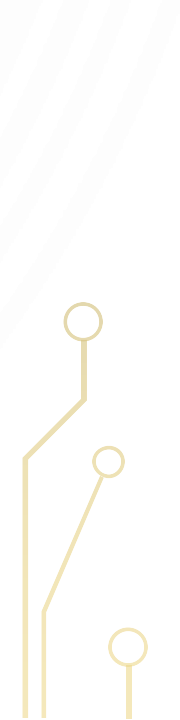
**SIDLIT 2014 (OF C2C )**

**JULY 31 – AUG. 1, 2014**



# PRESENTATION OVERVIEW



- Today, there are masses of texts being generated and shared publicly. There are microblogging Tweet streams and conversations; long-running blogs and wikis; open-ended text responses on large-scale surveys; digitally-released novels, and machine-generated texts (such as SciGen). In LMSes, there are numerous threads of student-generated conversations. Text has long been plumbed for meaning—based on context and so-called “close reading” by scholars. Of late, widely accessible computational tools have enabled text-mining. AutoMap and ORA NetScenes are tools created by CASOS (Center for Computational Analysis of Social and Organizational Systems) of Carnegie Mellon University that enable basic text-mining and the visualization of networked text in 2D and 3D formats. AutoMap, a text-mining tool, offers some basic methods for sentiment analysis, the extraction of ngrams, the definition of network text relationships, and other revelatory insights.
- 
- 

# WELCOME!

- Hi! Who are you? 😊
- What are your research areas of interest? What sorts of text datasets do you have access to that you might want to use for research?
- What have your experiences been with network-based text analysis (if any)? Any experiences with AutoMap? ORA-NetScenes?
- What would you like to learn during this session?





# A REVIEW OF TERMS

- **Text network analysis:** The study of connectivity between words and phrases in a text in order to identify core meanings in a text
- **AutoMap:** A software tool that enables the mapping of textual data based on Network Text Analysis
- **N-gram:** A contiguous string that serves as a unit in computational linguistics
- **ORA NetScenes:** A software tool that enables data extraction from social media platforms and the visualization of relational data in network graphs
- **DyNetML:** XML interchange language containing relational and network data

# A REVIEW OF TERMS (CONT.)

- **Text mining:** Extracting meaning from text
- **Text-level concepts:** Specific concepts (ideas) from the contents of the dataset
- **Higher-level concepts:** Generalized text-level concepts (ideas)
- **Semantics:** Meaning contained in words and phrases
- **Syntax:** Orderly rule-based structure of words and phrases to make meaning in a language
- **Anaphora:** A word which refers to a prior used word (to avoid repeating the initial term)

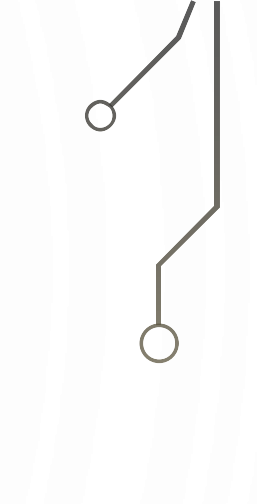
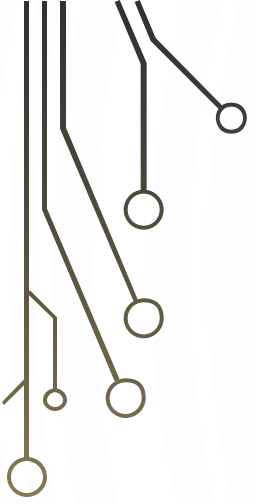


## A REVIEW OF TERMS (CONT.)

- **Scalability:** The ability to handle larger and larger amounts of data; the ability to “scale up” or “scale down” adaptably
- **Big data:**  $n = \text{all}$
- **Script:** A human-readable high-level programming language
- **Manual or hand coding:** The human extraction of nodes / codes through a close reading of text in a dataset (vs. automated coding or “autocoding” using automatic extractions or human-created thesauruses applied to text in a computer tool)

## A REVIEW OF TERMS (CONT.)

- **Data extraction:** The retrieval or downloading of data from a database
- **Meta-network:** An extracted network (from text corpuses) which is created either automatically or through a human-coded thesaurus (or a mix of both)
- **Graphical User Interface (GUI):** A screen-based user interface that enables people to interact with a computing system
- **Encapsulation:** The hiding of complexity within a simple interface (which may mask the actual functions of the software tool to users)



**Notes:** This presentation is really to provide an overview of some generally readily available capabilities. This is not to show necessarily “how” to achieve this in this short session.

Also, the presenter has only recently started experimenting with this software tool. This only shows one basic “use case” with many other software functionalities and use cases left unaddressed.

# MACHINE-BASED TEXT ANALYSIS







# WHY NETWORK TEXT ANALYSIS?

- **A Structured Way of Understanding the Main Concepts and Concept Interrelationships in Text Corpora:** The application of network science to texts, enabling fast captures of “gists” both within and between text corpora
- **Plenty of Source Material:** Broad prevalence of “big data” for analysis (archival data, digitized materials, social media contents, online prosumer-created contents, and others), which requires speed and scalability to reduce data to a manageable size, to ultimately capture gist and meaning
- **Scalability and Efficiencies:** The ability to harness computational resources to map texts and text corpora speedily and then to create data visualizations

# WHAT IS NETWORK-BASED TEXT ANALYSIS?

- Decomposition (and decontextualization) of language as symbols and code (semantics / meaning and syntactics / non-meaning structures / parts of speech tagging)
- Application of a “bag of words” (multiset) approach
- Use of a “window” to look at word proximity as an indicator of relatedness and meaning (semantics); frequency counts of ngrams (words, symbols, phrases); clustering; word-pair linkages; adjacency matrices, and other statistical approaches
- Frequency as an indicator of importance and focus of the text corpuses
- A form of text mining through systematic and quantitative analysis of texts
  - Lexical link analysis is a type



# WHAT IS NETWORK-BASED TEXT ANALYSIS? (CONT.)

- Scalable with computational machine support for large-scale textual data
- Data reduction through stopwords / delete words lists; generalizations of concepts to sentiments or other coding (machine-encoding and human-encoding)
- Human-driven process, with researchers having to apply their knowledge of the field
  - Analysis based on statistical analysis, graph analysis, and close reading of some sample documents...along with domain expertise



# NETWORK-BASED TEXT ANALYSIS GRAPH VISUALIZATIONS

## UNDERLYING STATISTICS

- Matrix analysis
- Windowing (proximity measures)
- Centrality-betweenness measures
- Centrality-eigenvector
- Clique (subnetwork) count
- Geodesic distance,
- Probability, and others

## GRAPH VISUALIZATIONS

- Meta-networks depicted as node-link diagrams (words and phrases as actors / entities; relatedness as links / edges)
- Interdependence of nodes to make meaning in the texts
- Variety of graph layout algorithms based on algebra
- 2D or 3D



# BASIC TENETS OF NETWORK TEXT ANALYSIS

- Networks of semantically linked concepts may be created to summarize some aspects of the contents of a text or text corpus (to serve as mental maps)
- Quantitative methods have a role in complementing qualitative analysis and other more traditional types of exploratory text analysis (critique)
- Enables textual-visual summarizations of texts and text corpora
- May be applied to large-scale text corpora



# A SAMPLER OF SOME VISUALIZATIONS

## Sample 1: Concept List Viewer

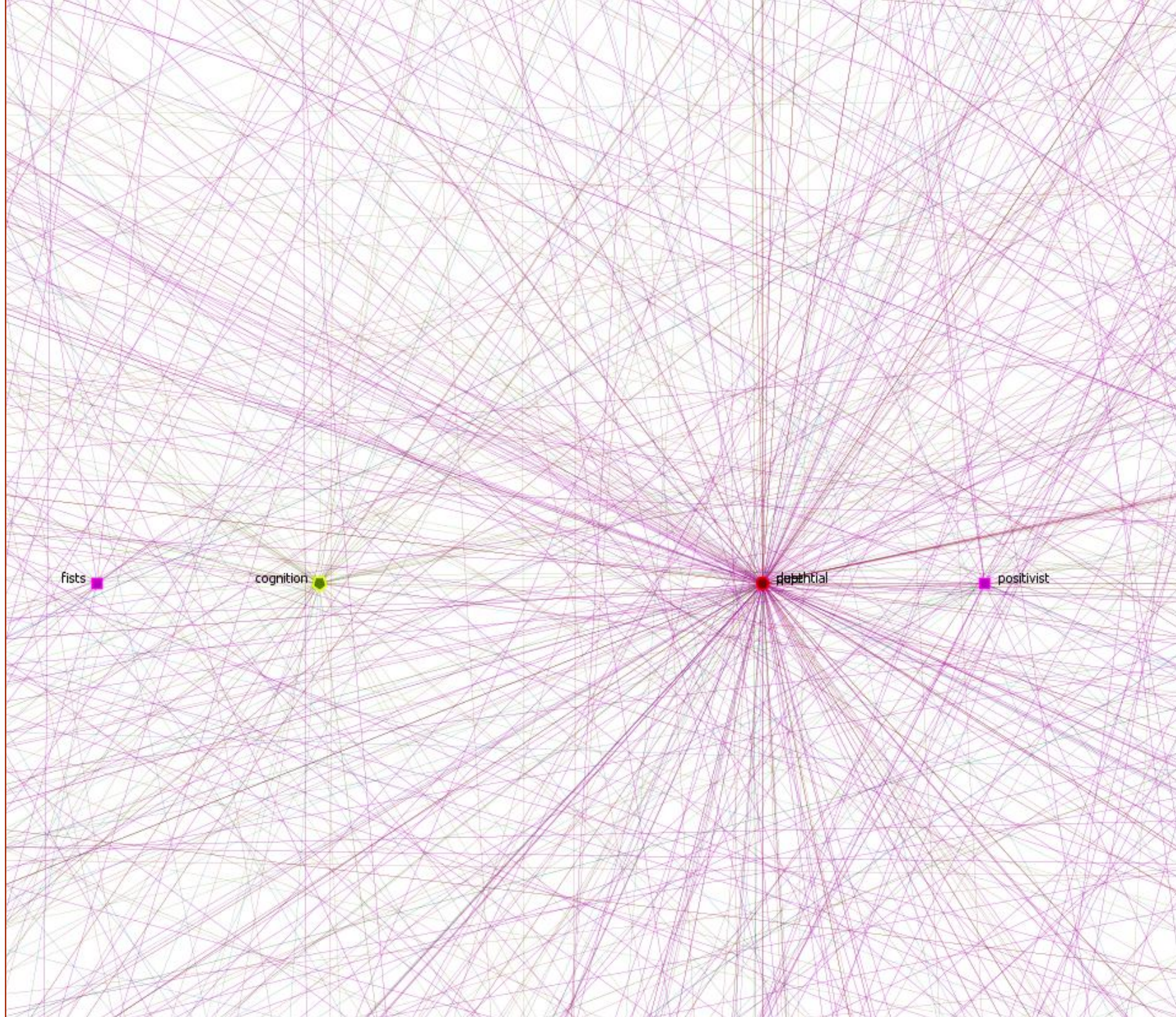
Concept List Viewer

File Edit Procedures Help

C:\Users\shalin.USERS\Desktop\2014SIDLITSlideshows\ConceptList2\union\union.csv

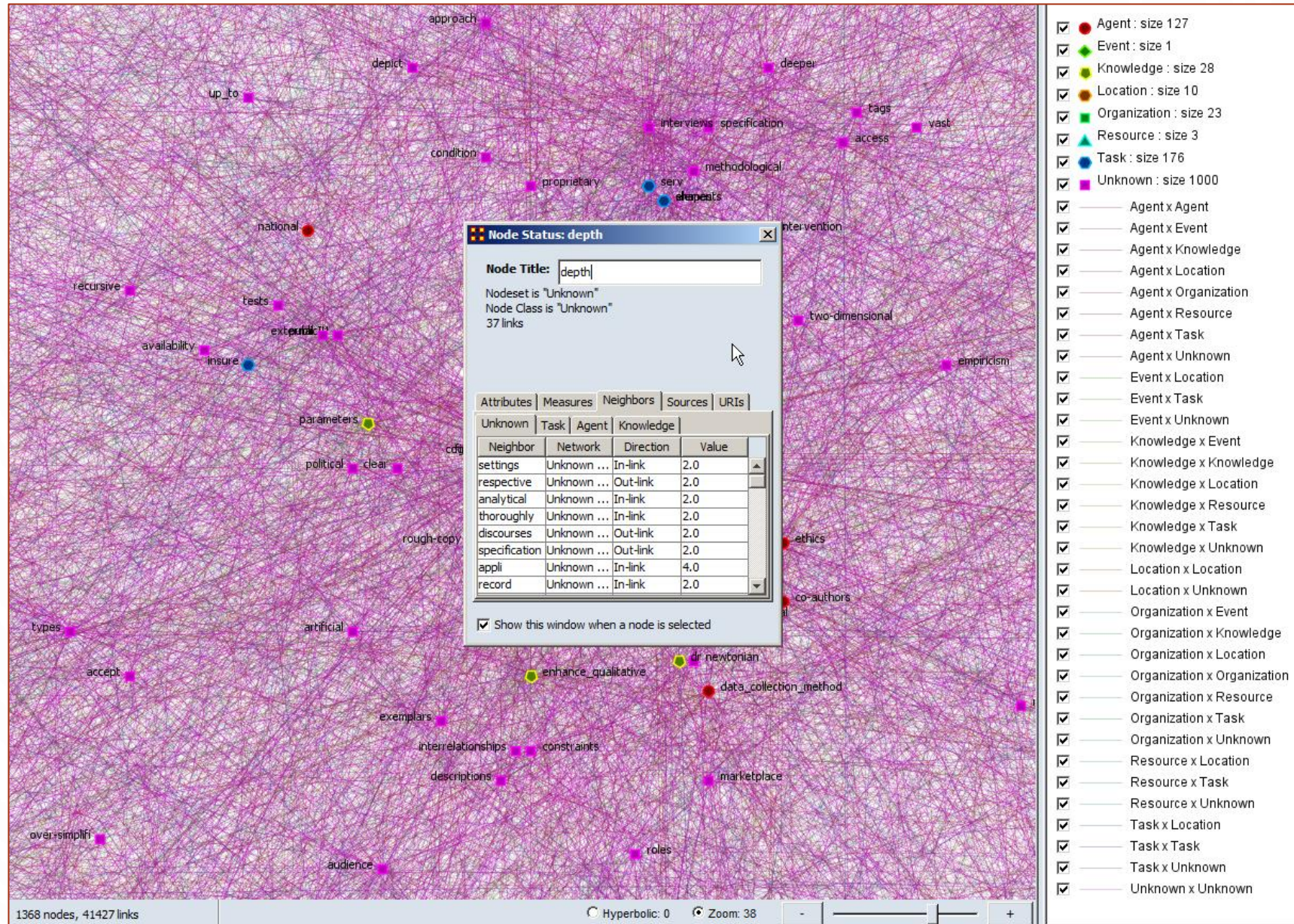
Select	concept	frequency	relative_frequency across texts	relative_percentage across texts	tf-idf	Action	Evaluation	Potency	gram_type	number_of_texts	pos
<input type="checkbox"/>	analyzing	1	Infinity	0.0	0.0				single	1.0	VBG
<input type="checkbox"/>	Grounded	2	Infinity	0.0	0.0				single	1.0	NNP
<input type="checkbox"/>	extractions	2	Infinity	0.0	0.0				single	1.0	NNS VBN
<input type="checkbox"/>	answered	1	Infinity	0.0	0.0				single	1.0	VBN
<input type="checkbox"/>	formats	1	Infinity	0.0	0.0				single	1.0	IN
<input type="checkbox"/>	metaphysics	1	Infinity	0.0	0.0				single	1.0	DT
<input type="checkbox"/>	spatialized	2	Infinity	0.0	0.0				single	1.0	DT IN
<input type="checkbox"/>	role	2	Infinity	0.0	0.0				single	1.0	NN
<input type="checkbox"/>	Overview	2	Infinity	0.0	0.0				single	1.0	DT NN
<input type="checkbox"/>	right	3	Infinity	0.0	0.0				single	1.0	JJ RB
<input type="checkbox"/>	FINAL	1	Infinity	0.0	0.0				single	1.0	IN
<input type="checkbox"/>	geographies	1	Infinity	0.0	0.0				single	1.0	NN
<input type="checkbox"/>	Among	1	Infinity	0.0	0.0				single	1.0	IN
<input type="checkbox"/>	Hart	2	Infinity	0.0	0.0				single	1.0	NNP
<input type="checkbox"/>	technology-augm...	1	Infinity	0.0	0.0				single	1.0	NN
<input type="checkbox"/>	evolution	1	Infinity	0.0	0.0				single	1.0	NN
<input type="checkbox"/>	few	5	Infinity	0.0	0.0				single	1.0	JJ
<input type="checkbox"/>	Platforms	5	Infinity	0.0	0.0				single	1.0	DT NNP
<input type="checkbox"/>	User-Created	2	Infinity	0.0	0.0				single	1.0	DT
<input type="checkbox"/>	3	9	Infinity	0.0	0.0				single	1.0	CD LS
<input type="checkbox"/>	1	11	Infinity	0.0	0.0				single	1.0	CD LS
<input type="checkbox"/>	2	13	Infinity	0.0	0.0				single	1.0	CD LS
<input type="checkbox"/>	7	1	Infinity	0.0	0.0				single	1.0	CD
<input type="checkbox"/>	0	1	Infinity	0.0	0.0				single	1.0	CD
<input type="checkbox"/>	5	5	Infinity	0.0	0.0				single	1.0	CD
<input type="checkbox"/>	6	5	Infinity	0.0	0.0				single	1.0	CD
<input type="checkbox"/>	Acknowledgments	1	Infinity	0.0	0.0				single	1.0	VBN
<input type="checkbox"/>	4	8	Infinity	0.0	0.0				single	1.0	CD LS
<input type="checkbox"/>	8	1	Infinity	0.0	0.0				single	1.0	CD
<input type="checkbox"/>	quality	5	Infinity	0.0	0.0				single	1.0	NN
<input type="checkbox"/>	dynamic	1	Infinity	0.0	0.0				single	1.0	JJ
<input type="checkbox"/>	D	2	Infinity	0.0	0.0				single	1.0	NN NNP
<input type="checkbox"/>	behavior	1	Infinity	0.0	0.0				single	1.0	NN

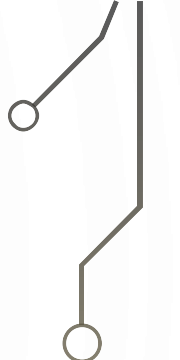
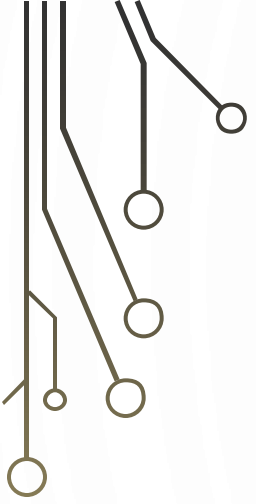
**Sample 2:  
Deeply Zoomed-  
in View on a  
Particular Node  
in a Meta-  
Network**



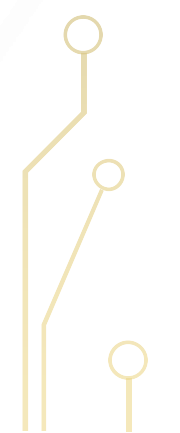
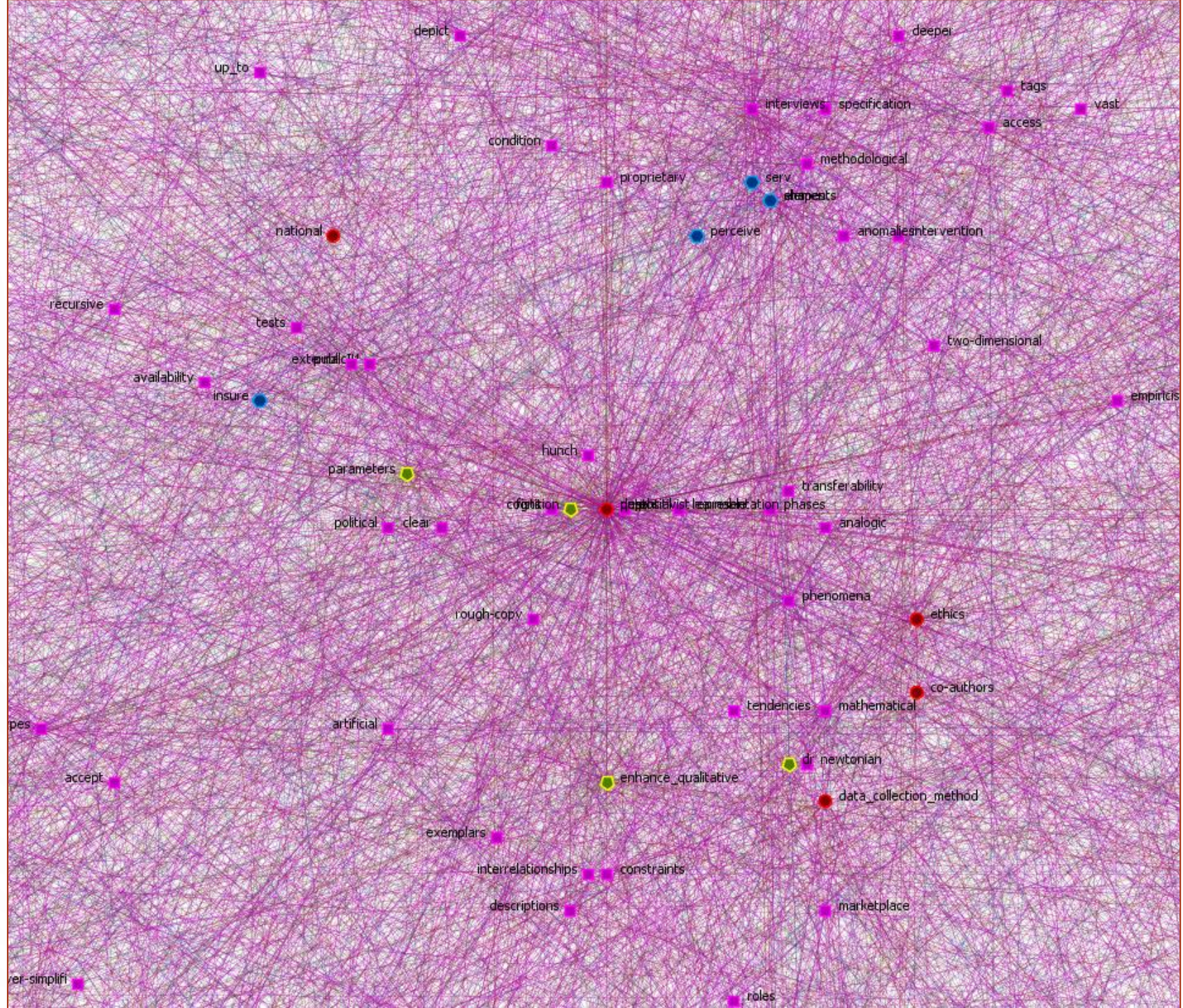


## Sample 3: Node-Level Details with a Click on the Node

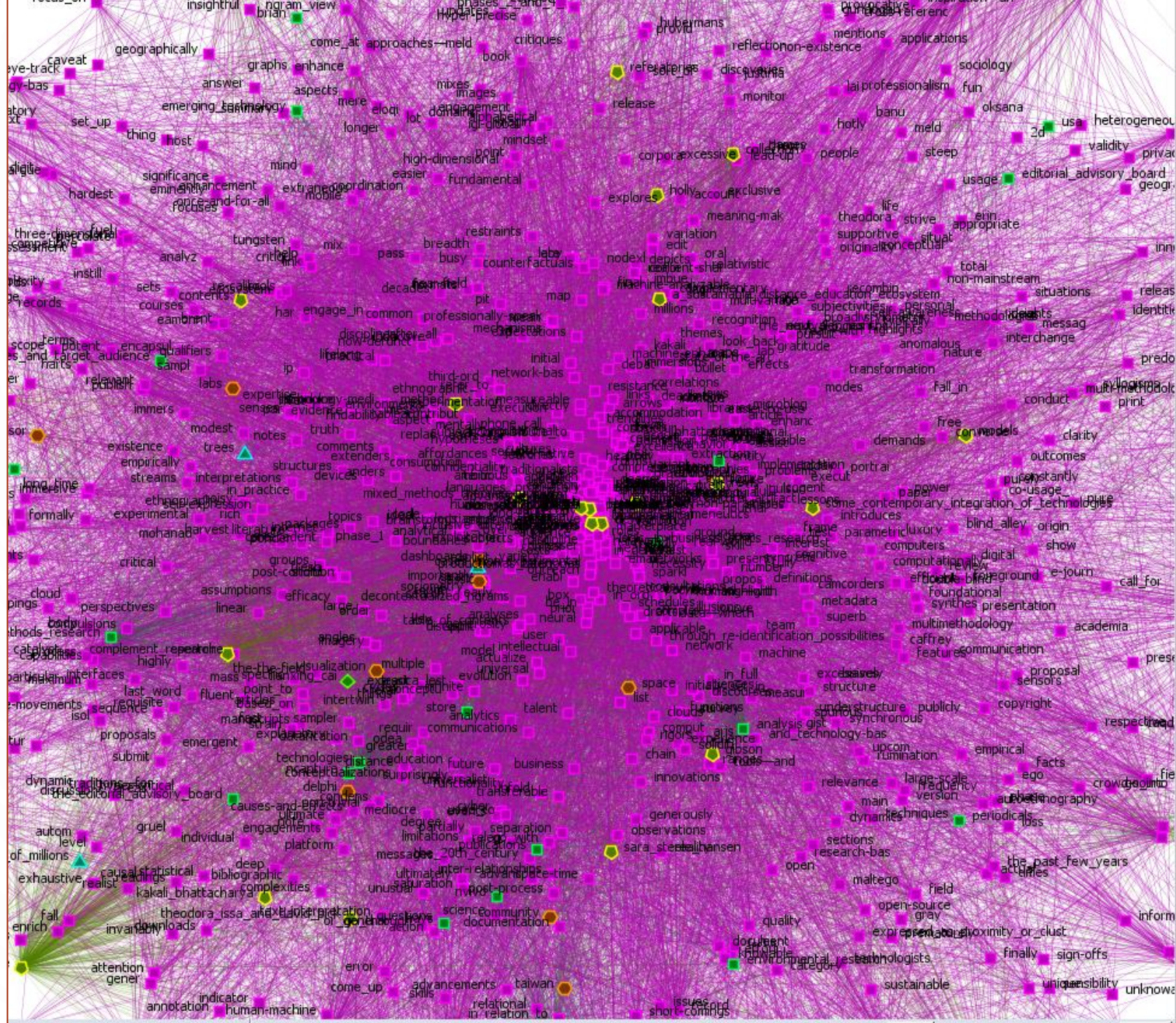




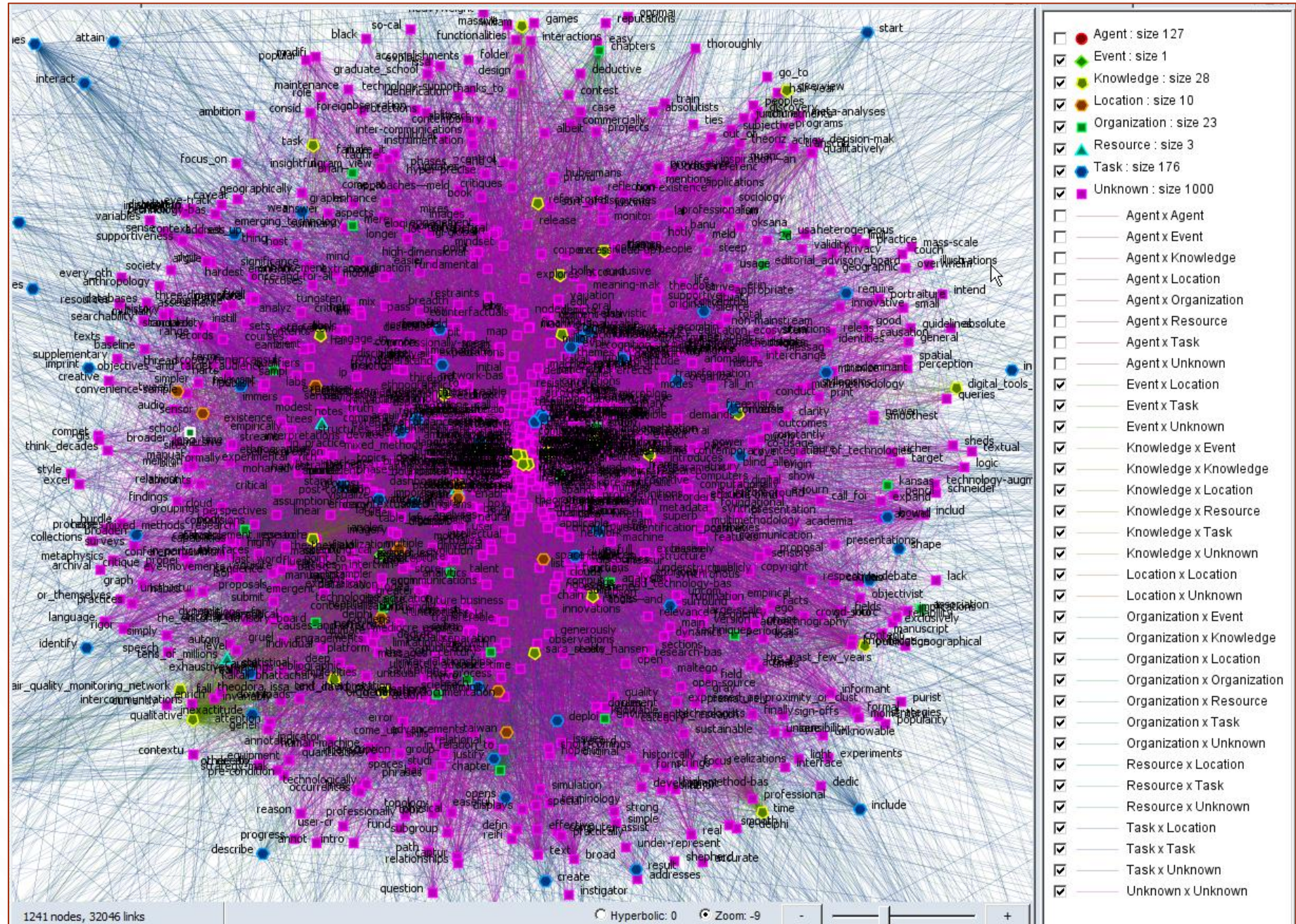
## Sample 4: Zooming Out



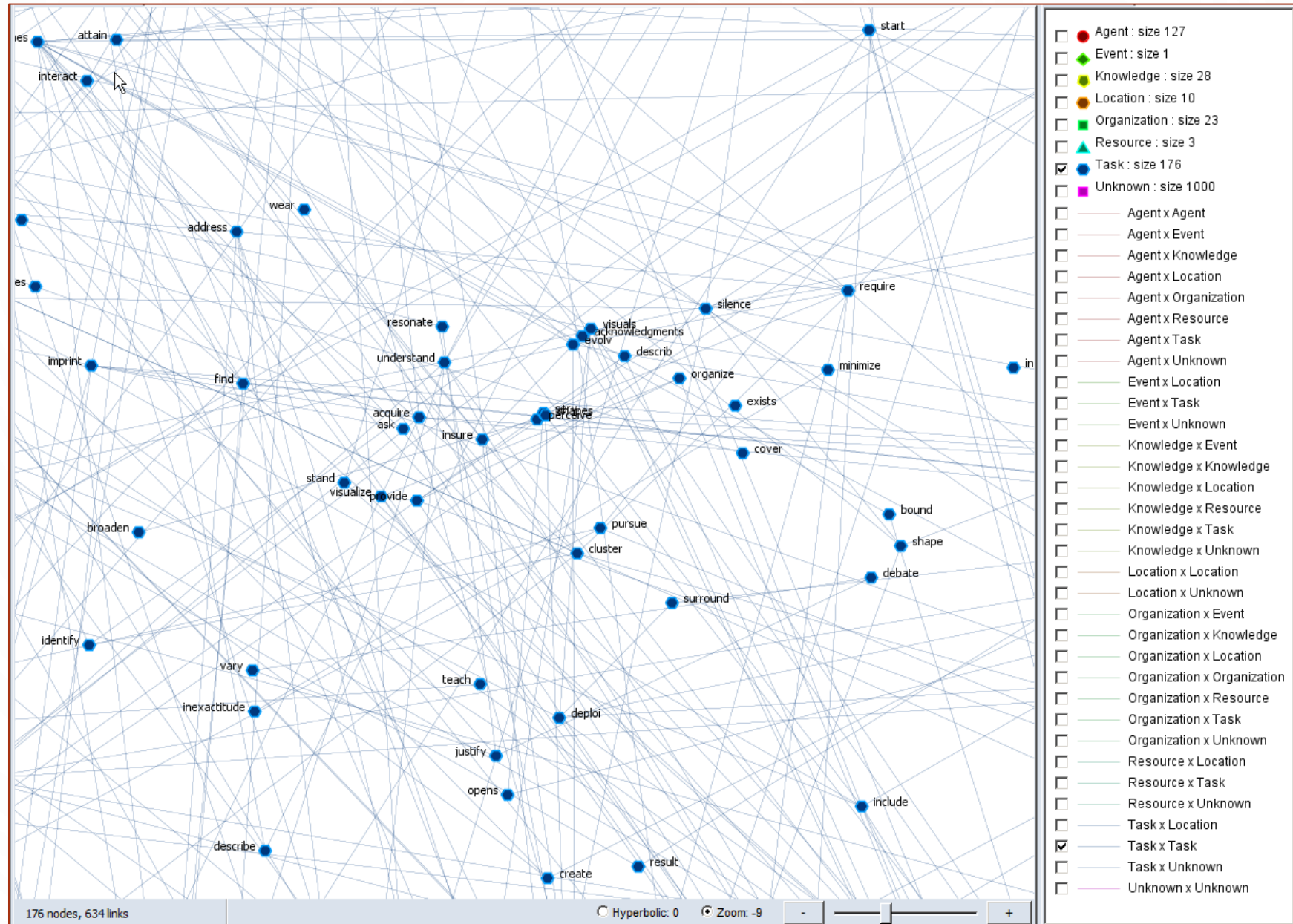
# Sample 5: Global Network View



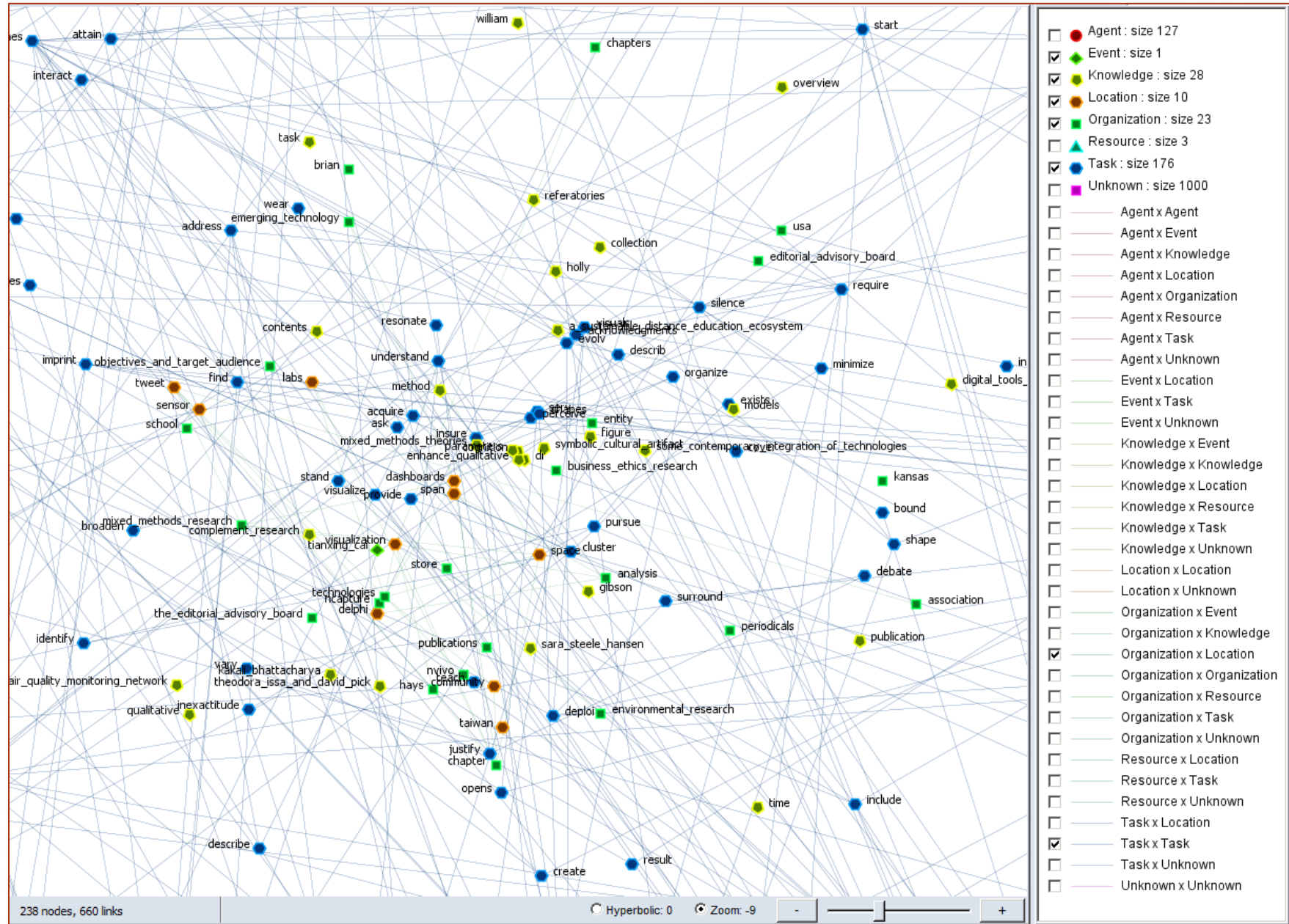
## Sample 6: Task Highlighted, Node Classes in the Right Column



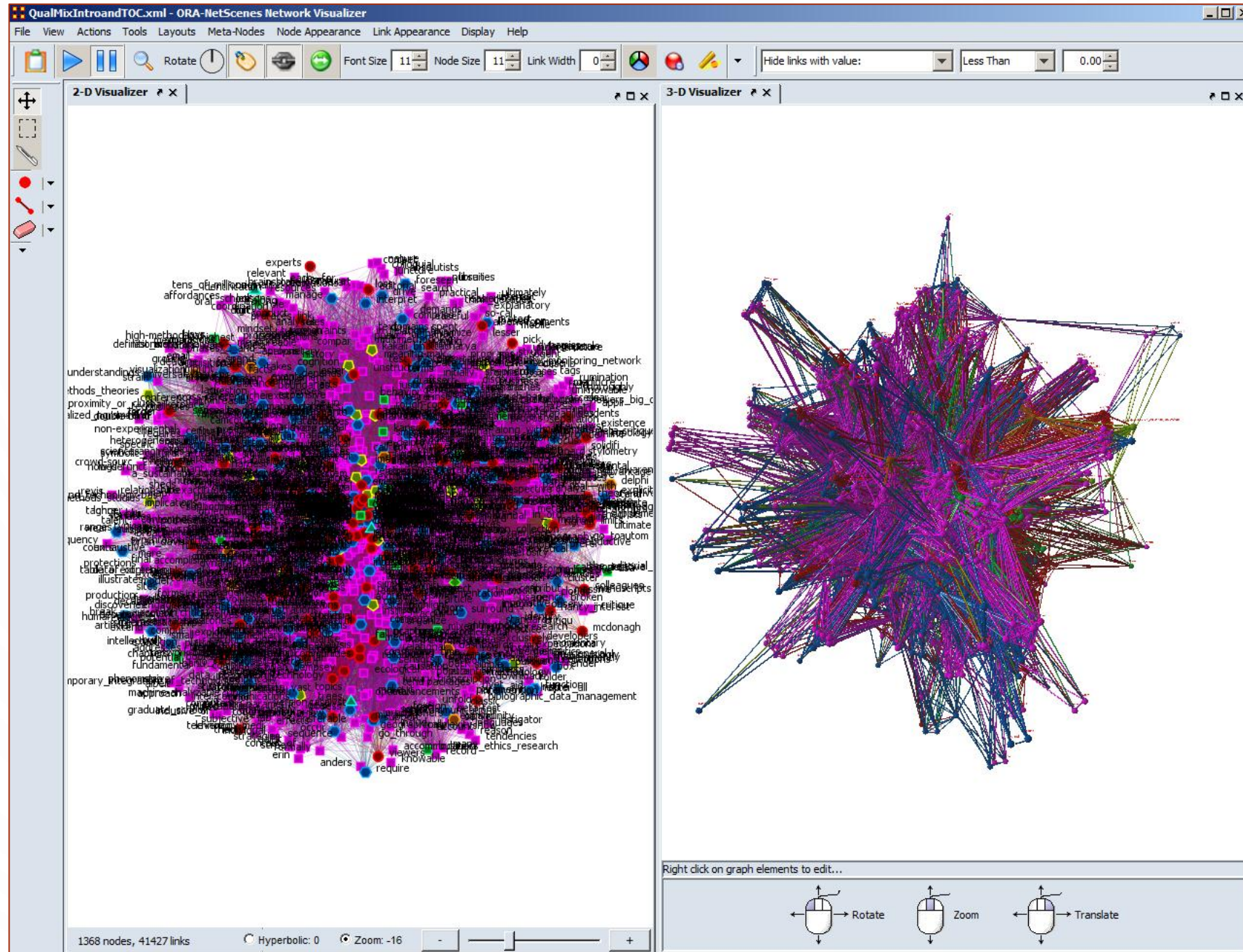
## Sample 7: Thinning the Network to Just Tasks (by Unchecking Boxes to the Node Classes)



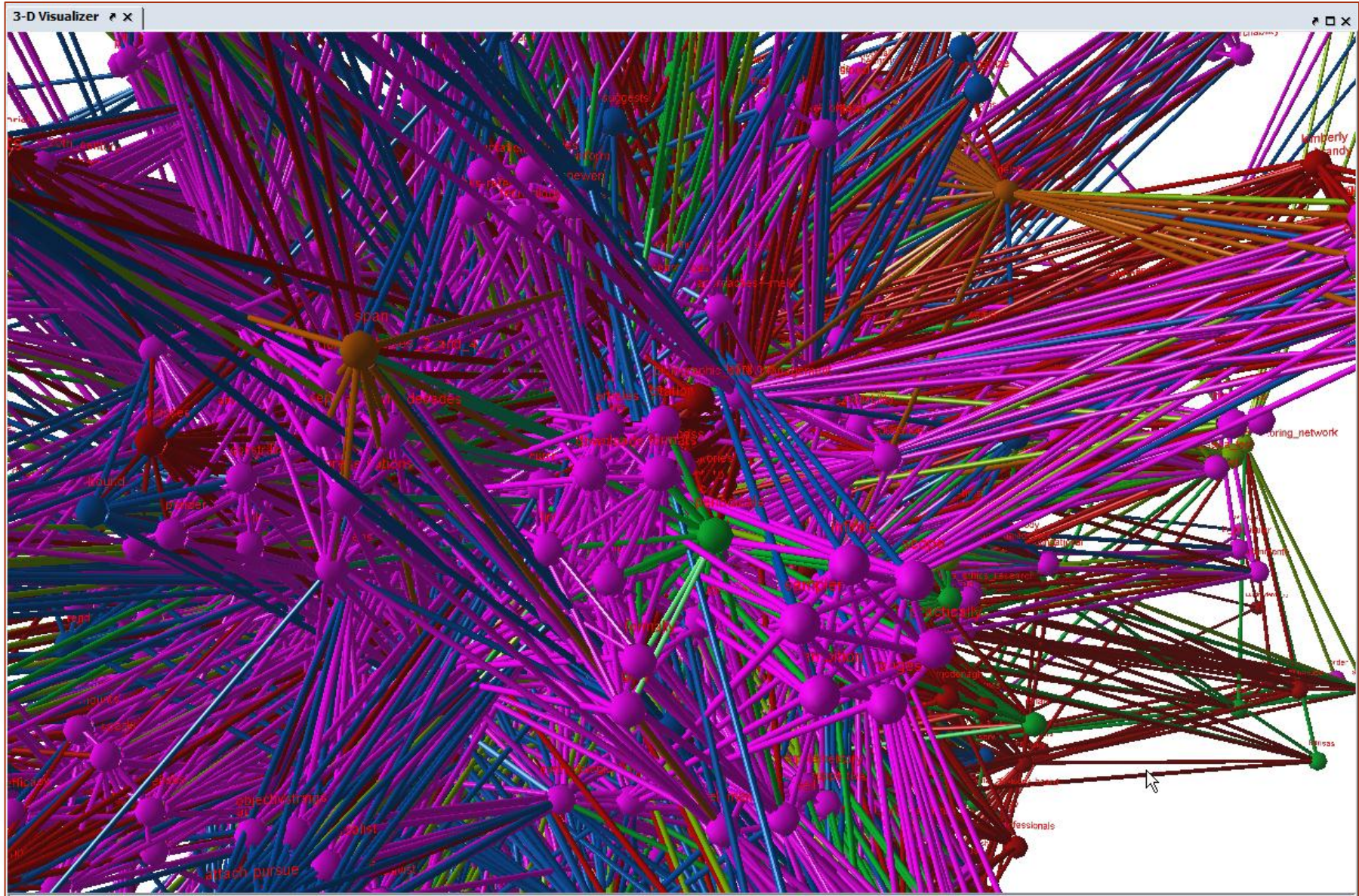
## Sample 8: Slow Re-introduction of Classes of Nodes for Analysis



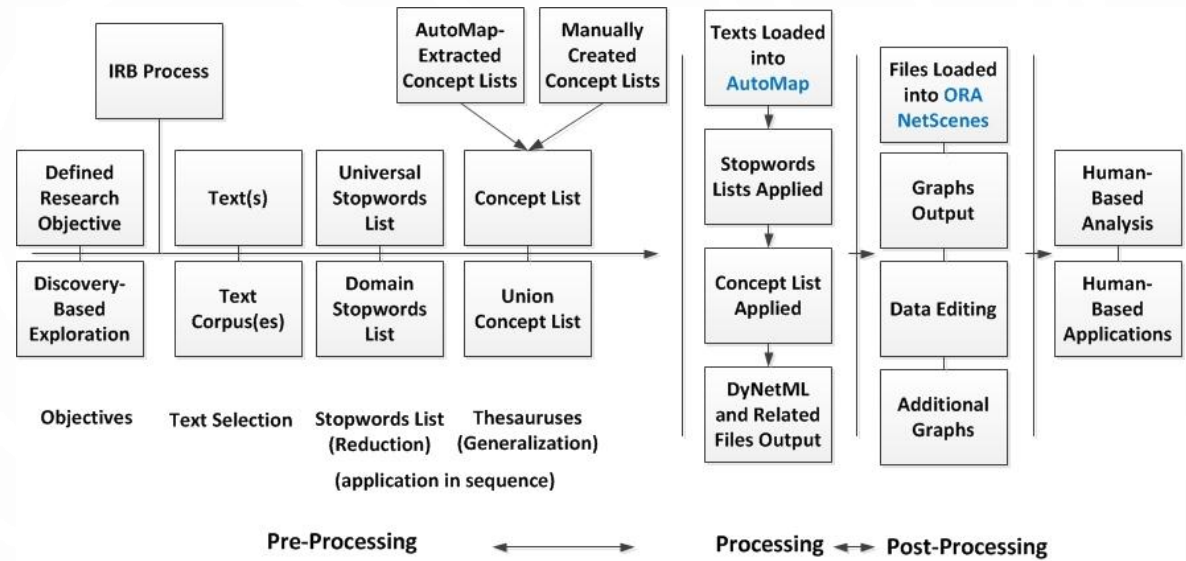
**Sample 9: 2D  
and 3D  
Visualizations  
(left and right  
respectively)**



**Sample 10:  
Zoomed In 3D  
Views**



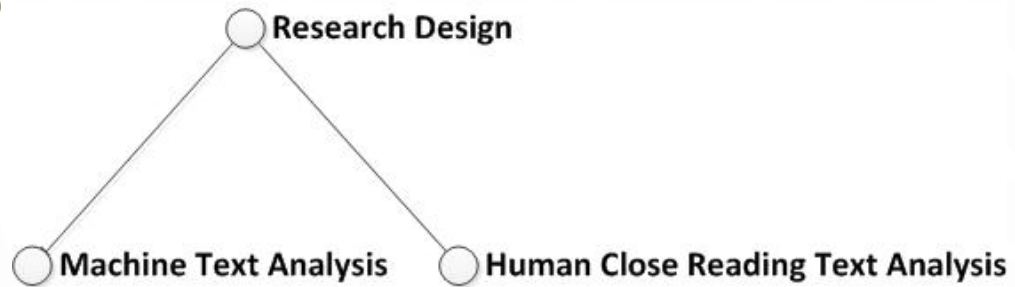




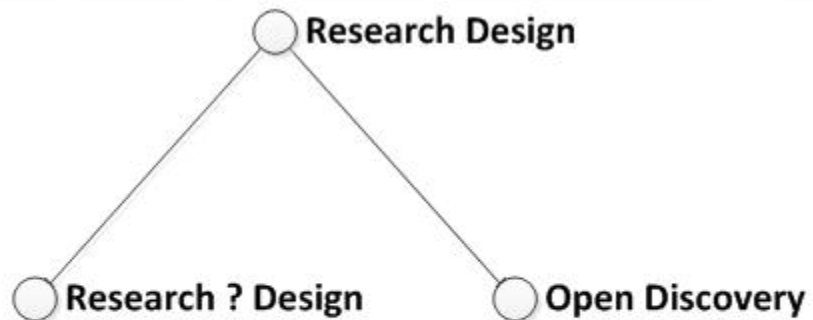
AutoMap Autocoding and ORA NetScenes Visualization for Network-Based Text Analysis  
(One Conceptualization of the Process)

# THE GENERAL WORKFLOW

# RESEARCH DESIGN



Not Either-Or, but Both-And



*A priori* Research Question Design  
or Open Discovery Research? Not Either-Or but Both?

- Research objective(s); definition of research questions and data queries
- Identification and acquisition of textual datasets (access matters)
- Openness to discovery



# CLASSIC INSTITUTIONAL REVIEW BOARD (IRB) ROLE IN REVIEW OF HUMAN SUBJECTS RESEARCH

- “Common Rule” of IRB oversight (particularly in research related to health and other sensitive issues)
- Multi-domain and trained IRB team analyzes all IRB proposals: the entire research study plan, the research rationale, the research team (and their credentialing and knowledge of research ethics and practices), recruitment strategies for participants (and the fair treatment of people), participant recruitment materials, informed consent forms, the maintenance of data throughout and after the study
- Caution with vulnerable populations, potential harm, any deception, and other aspects

# INSTITUTIONAL REVIEW BOARD (IRB) ROLE IN THE ANALYSIS OF SOCIAL MEDIA CONTENTS?

## YES TO IRB GUIDANCE

- Potential intrusiveness of social media data when analyzed using a variety of mining tools (structure mining, text mining, and others) beyond general public's knowledge
- Ease of re-identification of individuals with a few data points (with resulting privacy violations)

## NO TO IRB GUIDANCE

- The data is already public and broadly available
- Infeasible to contact all to get permissions to use their information
- Already have rights to the data within the EULAs of the social media platforms

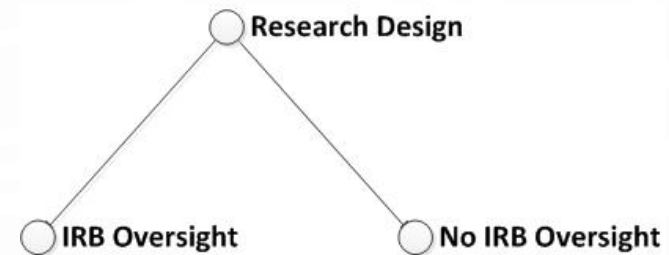
# INSTITUTIONAL REVIEW BOARD (IRB) ROLE IN THE ANALYSIS OF SOCIAL MEDIA CONTENTS? (CONT.)

## YES TO IRB GUIDANCE

- Control for a researcher spontaneously going off the approved plan and interacting with the participants (continuing oversight of IRB)
- IRB seeing potential unintended consequences
- Some legal cover

## NO TO IRB GUIDANCE

- If data properly handled, no one will be identified to the individual person
- IRB process requires effort and time



Human Subjects Review for Social Media Data?  
Analysis of Text Corporuses?

# SOME TENETS IN THE FAIR TREATMENT OF PEOPLE IN RESEARCH

## PROPER TREATMENT

- Sufficiently informed?
- Protected from harm?
- Ultimately benefitting from the research?
- Sufficient opt-out?

## IMPROPER TREATMENT

- Unaware and uninformed about the data use?
- Exposed to potential harm (risk) or actual harm
- Left out from actual benefits of the research
- “Outed” or identified / re-identified against their awareness or will?
- Unable to opt-out

# IDENTIFICATION AND ACQUISITION OF DATA FOR TEXT CORPUSES

- Some types of text-based datasets
  - Raw from-world datasets (microblogging data from #hashtag conversations, Tweetstreams, wiki data)
  - Formal and semi-formal datasets from periodicals (vetted by journalists and editors)
  - Processed datasets from other studies
  - Large-scale datasets made available to the public for study, and others

# QUALITY ISSUES WITH TEXTUAL DATASETS

- Factual vetting or not?
- Translated or not (and the possible unintended introduction of error)?
- Provenance?
- Digitized paper copies (with potential introduced errors from the OCR scanning)?
- Transcription of audio or video (with potential introduced errors from the transcription)? Etc.
- Comprehensiveness of the text corpuses?



# CREATION AND SELECTION OF TEXT CORPUS(ES)

## DATA STRUCTURED TEXT

- Labeled text in databases and spreadsheets

## DATA UNSTRUCTURED TEXT

- May relate to textual contents from social media platforms (like #hashtag conversations around particular issues, #eventgraphs, Tweet streams, blogs, websites, or other types of textual information)
- Manuscripts or other collections of text

# DATA PRE-PROCESSING

- Combinations of various texts into one...or into separate but related text corpuses that will be processed together (.txt)
  - Effectively removes images and formatting
- Creation of a “stopwords” / “delete” lists
  - Removal of “noise” (conjunctions, articles, proper nouns, relative pronouns, and words with syntactic but often not semantic value) from the data to leave the semantic data; removal of punctuation; removal of repeated data
- Use of the “remainder data”

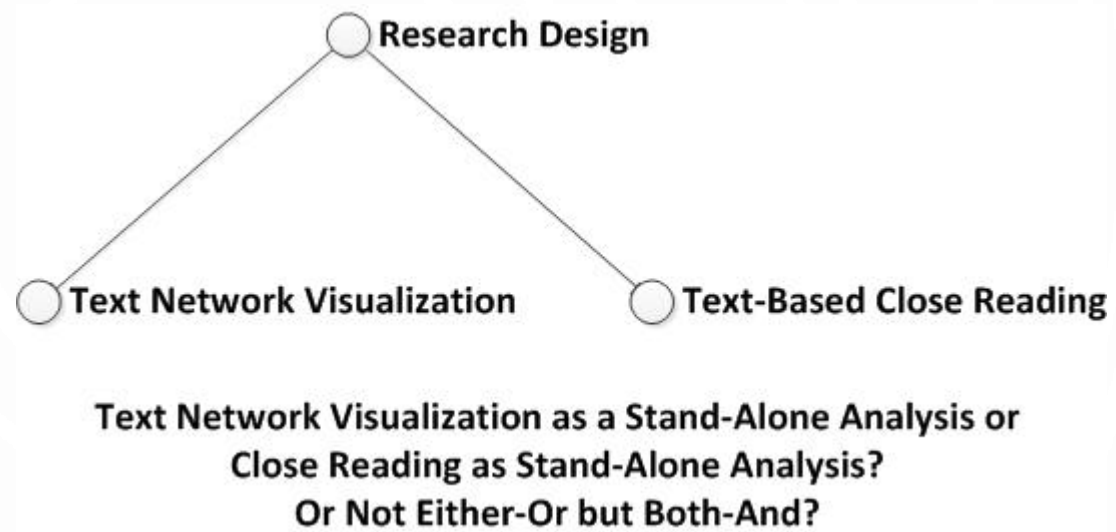
# DATA CLEANING (AKA SCRUBBING, CLEANSING)

- Application of the universal stopwords list, then the domain stopwords list
- Getting data into the optimal state on which to run queries without losing critical data (and without limiting the findings from the tool)
  - Elimination of corrupt records (which may crash the software)
  - Elimination of duplicate data (which can skew results)
  - Filling in gaps in the data (which may introduce error)
  - Omitting inaccurate records (which may introduce noise)
- Often done per run...without permanent effects on the textual corporuses

# DATA CODING

- A nontrivial and effortful process of structuring the data extraction through the uses of “thesauruses” (to reduce the data using delete lists and then to identify out important concepts and generalize the contents into this structure) using AutoMap
- Creation of generalization thesauruses by machine or by hand or by both
- Generalization thesauruses serve as a kind of code-book for AutoMap
- Application of the Concept List / Union Concept List
- Data run
- Data visualization of meta-networks
  - The inherent ambiguity of language (synonymy / synonymous; polysemy or multi-meaningness / openness to interpretation)

# HUMAN ANALYSIS





# STATISTICAL ANALYSES

- Network structure (degree distribution, density, linkages, geodesic structure)
- Edges (weight)
- Paths (degrees of separation between words)
- Nodes (degree, position of words or phrases in the text document or text corpuses; frequency counts)
- Hubs or regions or topics of interest
- Subgraph densities (interconnected semantic clusters or fields)
- Clusters (connected components based on clustering coefficient, filtering, to show deeply related ideas or concepts)

# GRAPH VISUALIZATION ANALYSES

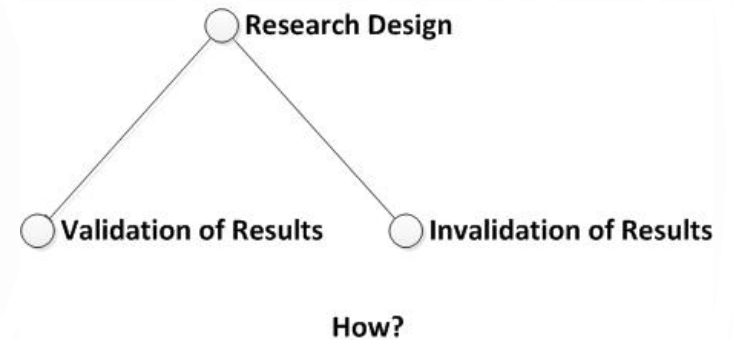
## TEXT-BASED GRAPHS

- Word hubs with linkages and branching off
- Linked words and phrases
- The identification of themes
- Knowledge structures
- Center-periphery dynamics of texts

## OTHER RELATED VISUALIZATIONS

- Word-tree diagrams and linked ideas
- Word clouds

# “PRESSURE TESTING” THE RESULTS: VALIDATION, INVALIDATION



## Historically...

- Computational models have been compared against human experts evaluations who've read the close datasets and created ontologies or other “maps” of the data
- Datasets have been run on various software tools to compare outputs
- Comparability to the researchers' close readings of a sampling of the information; comparability to the researchers' domain expertise and knowledge of the field



# “PRESSURE TESTING” THE RESULTS: VALIDATION, INVALIDATION (CONT.)

## Quality Sample Question List

- How comprehensive or inclusive is the extracted meta-network? What data is not appearing (which should appear)? Any anomalies? Are these anomalies informative?
- How apparently accurate is the frequency count for main terms or phrases or agents? Any insights (not knowable otherwise)?
- How (in)accurate are the relationships identified by the visualization?



# “PRESSURE TESTING” THE RESULTS: VALIDATION, INVALIDATION (CONT.)

- How coherent or intelligible are the visualizations?
- What new insights have been raised by the network-based text analysis?
- How useful are these insights or leads in follow-up research through other channels?
- From such validation / invalidation feedback, there may be changes to the research design, including the choice of datasets, the data scrubbing, the types of data visualizations, and the analytical techniques.

# CROSSING THE RUBICON

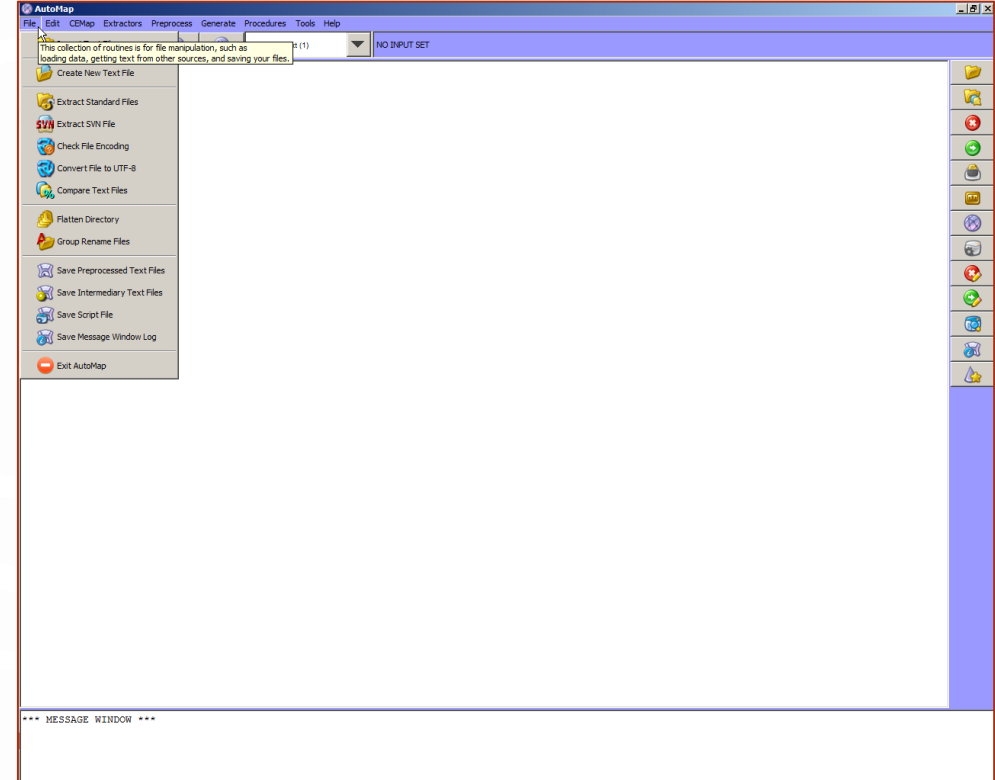
...AND INTO THE WEEDS



# AUTOMAP

BY CASOS (CENTER FOR COMPUTATIONAL ANALYSIS OF SOCIAL AND ORGANIZATIONAL SYSTEMS) AT  
CARNEGIE MELLON UNIVERSITY

AUTOMAP = (HUMAN-MACHINE) MANUAL AND AUTOMATED CODING -> TEXT MAP (GRAPH)



# A BRIEF HISTORY OF THE TOOL

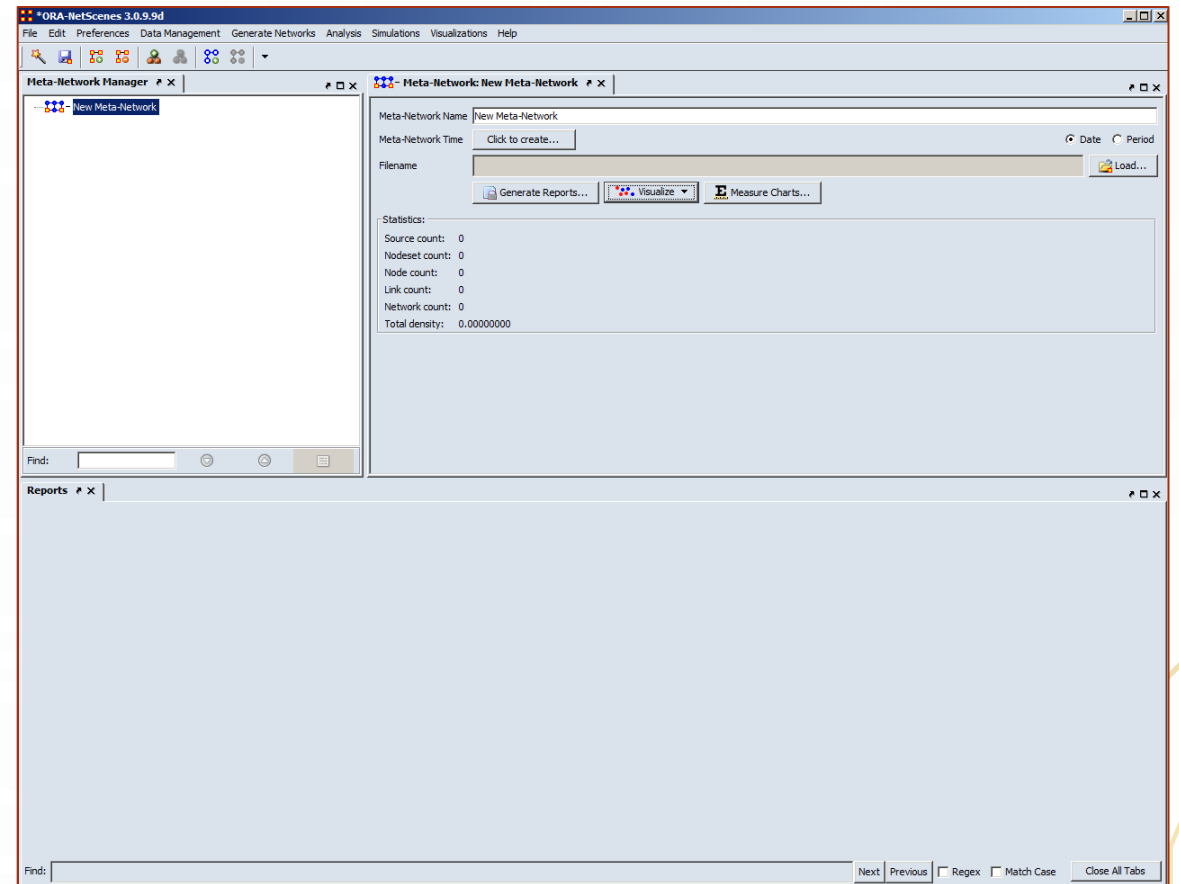
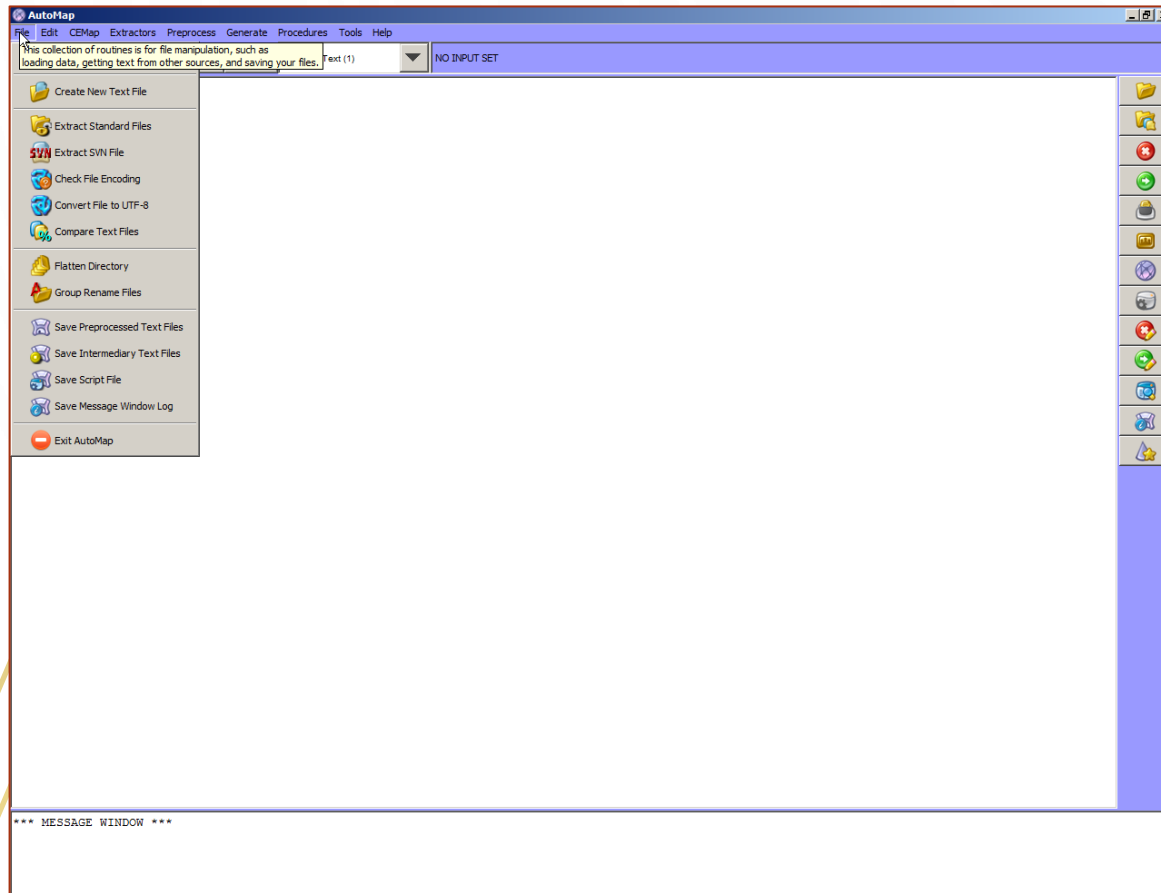
- Originated in [2001](#) (albeit with years of lead-up work and research)
  - Enables map analysis of texts (including meta-matrix analysis and sub-matrix analysis)
- Latest Version: [AutoMap 3.0.10.18](#) for Windows 32-bit and Windows 64-bit
- [AutoMap User's Guide 2013](#) (Carley, Columbus, & Landwehr, 2013)



# CONDITIONS FOR USE

- Research usage only; commercial usage possible through [Netanomics](#)
- Crediting required:
  - COPYRIGHT (c) 2001-2014 Kathleen M. Carley - Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research International (ISRI), School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue - Pittsburgh, PA 15213-3890 - ALL RIGHTS RESERVED.

# GRAPHICAL USER INTERFACES: AUTOMAP AND ORA NETSCENES





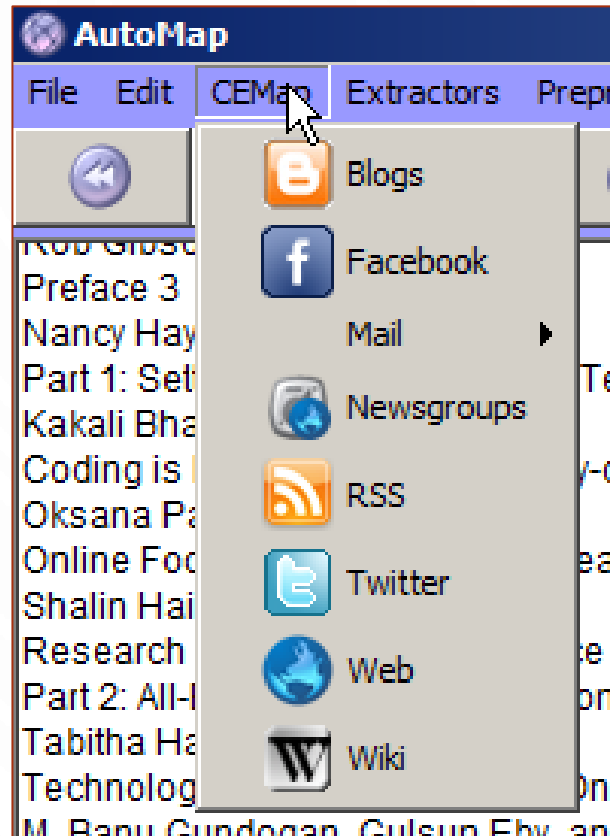
# TYPES OF DATA EXTRACTIONS FROM AUTOMAP FOR NETWORK MAPPING

1. “content (concepts, frequencies and meta-data such as sentence length);
2. semantic networks (concepts and relationships);
3. meta-networks (ontologically coded concepts and relationships—named entities and links);
4. sentiment and node attributes (attributes of named entities)” (Bigrigg, Carley, Kunkel, Diesner, Eisenberg, Chieffallo, & Columbus, 2010)



# DATA EXTRACTORS FROM SOCIAL MEDIA

- Blogs
- Facebook
- Email
- Newsgroups
- RSS feeds
- Twitter



- Web
- Wiki



# SIX TYPES OF THESAURUS FORMATS

1. Single column format for stopwords / delete lists
2. Two-column generalization format
3. Two-column meta-network format
4. Master format
5. Reduced format
6. Change format (Sangal, Carley, Altman, & Martin, 2012)

# TYPE 1: SINGLE COLUMN FORMAT (FOR STOPWORDS / DELETE LISTS)

- Column A1
- No column header
- List of words which should not appear in the dataset to be analyzed
- A “stopwords” or “delete” list (pronouns, prepositions, punctuation, verbs, articles, prepositions, and proper nouns, etc.)
- Data reduction

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER

Clipboard Font Alignment Number Styles Cells

Calibri 11 A A Wrap Text General

B I U Merge & Center \$ % , .00 .00

Conditional Formatting Format as Table Cell Styles Insert Delete

Q32 : X ✓ fx

	A	B	C	D	E	F	G	H	I	J	K
1	and										
2	or										
3	but										
4	for										
5	so										
6	yet										
7	the										
8	a										
9	an										
10	from										
11	to										
12	above										
13	below										
14	under										
15	over										
16	within										
17	without										
18	around										
19	with										
20	before										
21	after										
22	during										
23	is										
24	are										
25	was										
26	were										
27	be										
28	been										
29	being										
30	http:										
31	then										
32	which										
33	whichever										
34	who										
35	whomever										
36	that										

FILE HOME INSERT PAGE LAYOUT

Clipboard Font

D34 : X ✓ fx

	A	B	C	D
1	and			
2	or			
3	but			
4	for			
5	so			
6	yet			
7	the			
8	a			
9	an			
10	from			

## TYPE 2: TWO-COLUMN GENERALIZATION FORMAT

- No column header
- Changing text-level concepts to high-level concepts (moving from the specific to the general)
  - One example is a two-column list associating emails in Column A to names of the persons in Column B (with underscores used in lieu of spaces as in `firstname_lastname`)
  - Another example may be the conversion of specific dates (Column A) into larger categories like years (Column B)
- Generalization function / coding / data reduction

## TYPE 3: TWO-COLUMN META NETWORK FORMAT

- No column header
- High-level (generalization) concepts -> metaOntology (node class: agent, resource, knowledge, task, organization, location, event, action, belief, role, group)

TwoColumnMetaNetworkFormat - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER

Clipboard Font Alignment Number Styles Cells

P25

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	jane_doe	agent														
2	john_doe	agent														
3	seattle_w	location														
4	portland_	location														
5																
6																
7																

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER

Clipboard Font

C14

	A	B	C	D	E
1	jane_doe	agent			
2	john_doe	agent			
3	seattle_w	location			
4	portland_	location			
5					
6					

31  
32  
33  
34  
35  
36

Sheet1

READY

## TYPE 4: MASTER FORMAT

- The two-column format plus attribute information
- Column headers: `conceptFrom`; `conceptTo`; `metaOntology` (node class: `agent`, `resource`, `knowledge`, `task`, `organization`, `location`, `event`, `action`, `belief`, `role`, `group`); `metaName` (metatype) (generic or specific concept)
- Text from the mss in A (`conceptFrom`); text-level concept in B (`conceptTo`); node class in C (`metaOntology`); whether concept is generic or specific in D (`metaName`)
- Also known as Union Concept List, Concept List



MasterFormat - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER

Themes Colors Fonts Effects Margins Orientation Size Print Area Breaks Background Print Titles Width: Automatic Height: Automatic Scale: 100% Gridlines View Print Sheet Options Headings View Print Bring Forward Send Backward Selection Pane Align Group Rotate Arrange

Q35

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	conceptfrom	conceptto	metaontology	metaname												
2	<a href="mailto:haijes@gmail.com">haijes@gmail.com</a>	Shalin_Hai-Jew	agent	specific												
3																
4																
5																
6																
7																
8																
9																
10																
11																
12																
13																
14																
15																
16																
17																
18																
19																
20																
21																
22																
23																
24																
25																
26																
27																
28																
29																
30																
31																
32																
33																
34																
35																
36																

The image displays three overlapping windows of the XML Viewer application. The background window shows a tree view for a 'DynamicNetwork' with a 'MetaNetwork' containing various 'propertyIdentities' and 'sources'. A middle window shows a list of nodes, with the bottom ones expanded to show their labels: 'node [ bind ]', 'node [ binding ]', and 'node [ bing ]'. The foreground window shows a detailed view of a selected node, displaying its 'propertyIdentities' (such as 'relative\_frequency-within\_text' with a value of 0.0068701627) and a list of sub-nodes.

# CONCEPT LIST WITH PART OF SPEECH (POS) EXTRACTIONS (7,238 CONCEPTS, 1,911 UNIQUE CONCEPTS)

- Can select certain concepts and resave the list as a “delete list”

The screenshot shows the 'Concept List Viewer' application window. The main window displays a table with the following columns: Select, concept, pos, frequency, relative\_frequency\_within\_text, gram\_type, number\_of\_texts, Evaluation, Potency, and Action. The table lists various concepts such as 'Communication', 'Dashboards', 'D', 'David', 'Data', etc., with their respective parts of speech and frequencies.

A 'Properties' dialog box is overlaid on the table, displaying the following information:

- Total Concepts: 7238
- Unique Concepts: 1911

The dialog box has an 'OK' button and a close button (X).

## TYPE 5: REDUCED (OR REVIEW) FORMAT

- 7 columns
  - Column headers: FREQUENCY, CURRENT\_CONCEPT, NEW\_CONCEPT, CURRENT\_METAONTOLOGY, NEW\_METAONTOLOGY, CURRENT\_METATYPE, NEW\_METATYPE
  - Subsumes Types 2, 3, and 4
  - New\_Metaontology enables deletions of words, ignoring of words, and splitting of phrases
  - Enables the changing of the meta-network structure
- Also known as Review Format

ReducedorReviewFormat - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER

Clipboard Font Alignment Number Styles Cells Editing

Calibri 11 Wrap Text General

B I U Merge & Center \$ % .00 .00

Conditional Formatting Format as Table Cell Styles

AutoSum Fill Clear Sort & Filter Find & Filter Select

P15

1	FREQUEN	CURRENT_NEW_CON	CURRENT_NEW_MET	CURRENT_NEW_MET	TYPE
2	Full Name	full_name	agent	agent	generic generic
3	City, State	city_state	location	location	generic generic

ReducedorReviewFormat - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW DEVELOPER

Clipboard Font Alignment Number

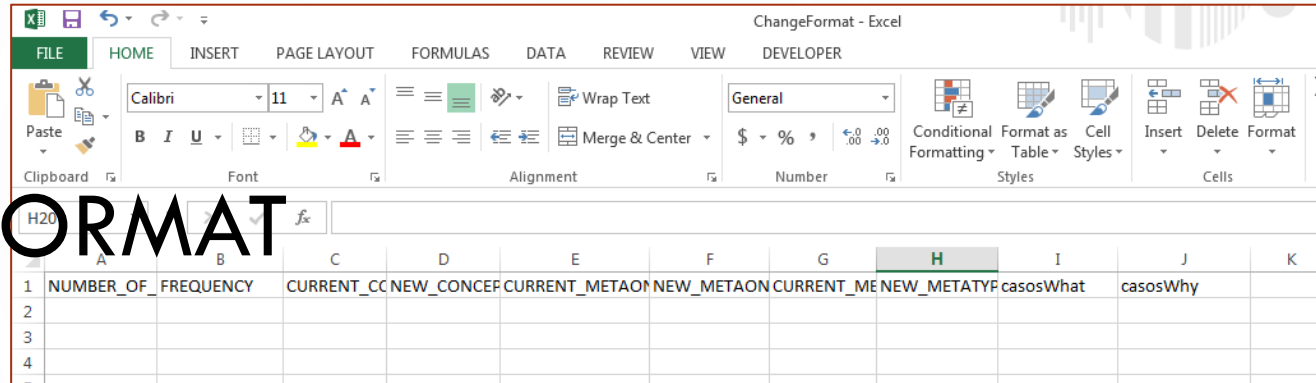
P15

1	FREQUEN	CURRENT_NEW_CON	CURRENT_NEW_MET	CURRENT_NEW_MET	TYPE
2	Full Name	full_name	agent	agent	generic generic
3	City, State	city_state	location	location	generic generic

Sheet1

READY 100%

# TYPE 6: CHANGE FORMAT



	A	B	C	D	E	F	G	H	I	J	K
1	NUMBER_OF	FREQUENCY	CURRENT_CC	NEW_CONCEP	CURRENT_METAON	NEW_METAON	CURRENT_ME	NEW_METATYP	casosWhat	casosWhy	
2											
3											
4											

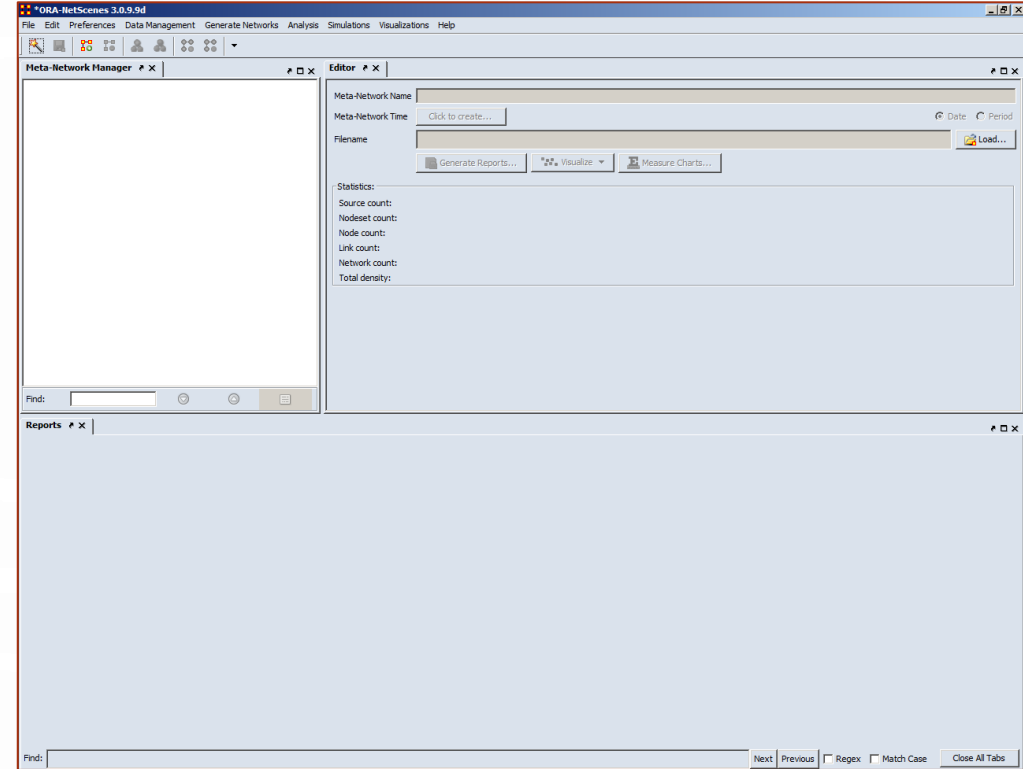
- Comprises 10 columns
- Subsumes all prior formats of thesauruses
- Column headers: NUMBER\_OF\_TEXTS, FREQUENCY, CURRENT\_CONCEPT, NEW\_CONCEPT, CURRENT\_METAONTOLOGY, NEW\_METAONTOLOGY, CURRENT\_METATYPE, NEW\_METATYPE, casosWhat, casosWhy
  - NUMBER\_OF\_TEXTS: The number of texts containing the concept
  - casosWhat: Name of the thesaurus file “from where the concept comes” (Sangal, Carley, Altman, & Martin, 2012, p. 9)
  - casosWhy: Explanation for mapping from Initial concept to the New concept
- Other columns may be added to allow for more complex graph visualizations and data queries

# REMixING OF ThESAURUSES

- AutoMap enables thesauruses to be combined and re-arranged through “split” and “merge” routines
- Universal thesauruses, for example, may be collated from a range of other thesauri
- Domain thesauri are specific to a project and contain only entities related to the project (and unique to the project); domain thesauri have precedence over universal thesauri (in terms of coding)

# ORA-NETSCENES

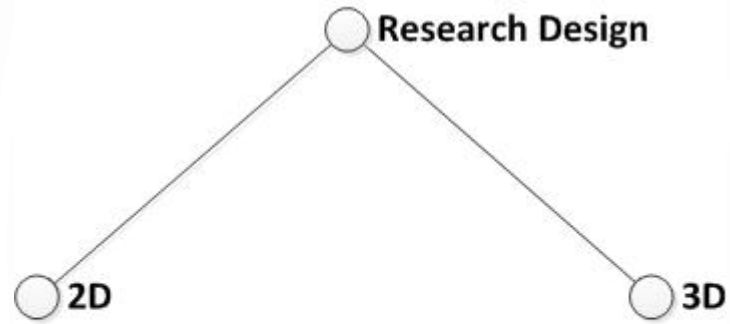
BY CASOS (CENTER FOR COMPUTATIONAL ANALYSIS OF SOCIAL AND ORGANIZATIONAL SYSTEMS) AT  
CARNEGIE MELLON UNIVERSITY





# A BRIEF HISTORY OF THE TOOL

- [ORA NetScenes 3.0.8.6](#) available for download in Windows 32-bit and 64-bit formats (from March 12, 2014)
- Created in 2001
- Available for non-commercial research use
- Crediting required:
  - COPYRIGHT (c) 2001-2014 Kathleen M. Carley - Center for Computational Analysis of Social and Organizational Systems (CASOS), Institute for Software Research International (ISRI), School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue - Pittsburgh, PA 15213-3890 - ALL RIGHTS RESERVED.



Two-Dimensional Visualizations or Three-Dimensional Visualizations, or Both?

# TWO-DIMENSIONAL OR THREE-DIMENSIONAL VISUALIZATIONS

## 2D

- Offers relative clarity of layout even with some fairly complex graphs
- Provides a fine sense of overview
- Works better as a print visual

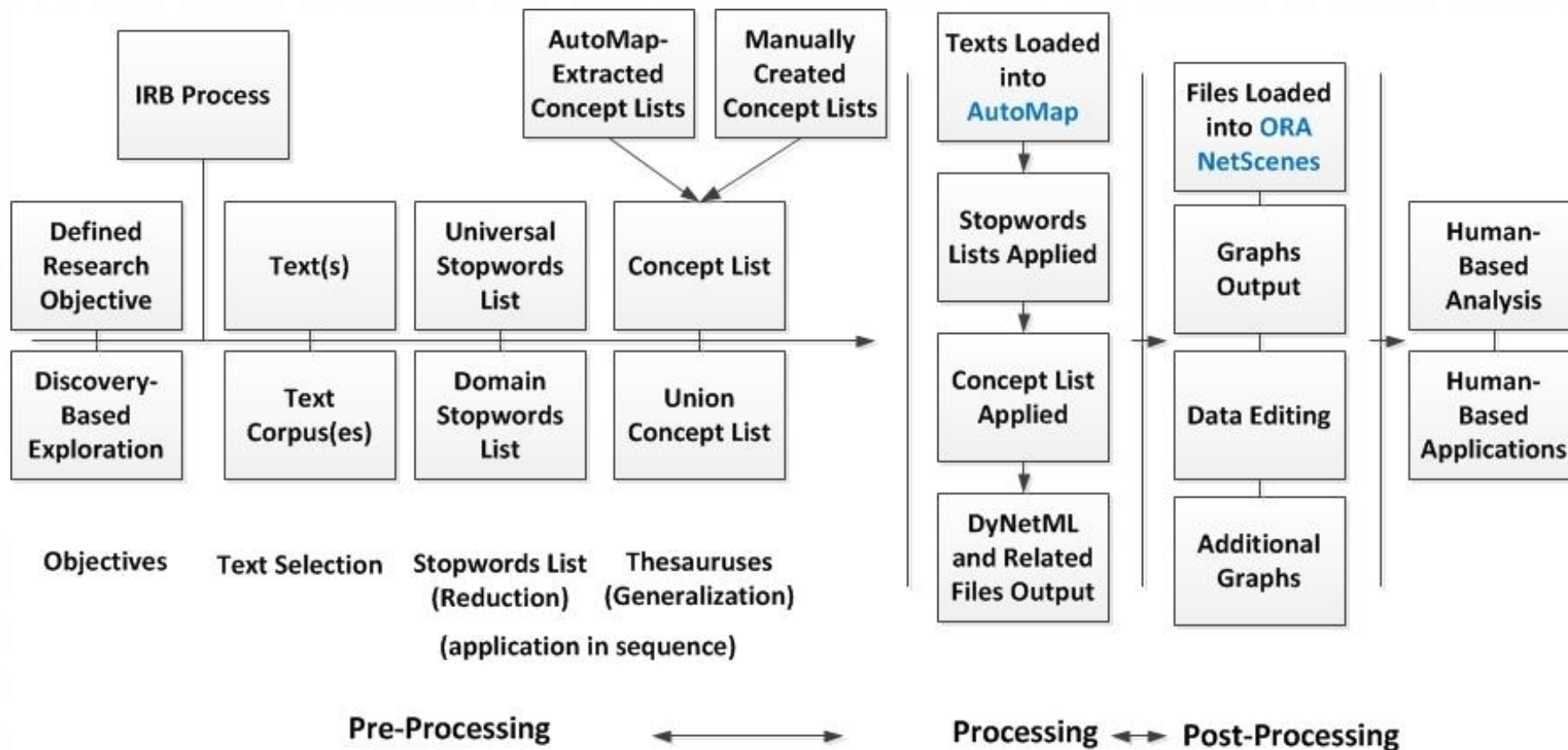
## 3D

- Can evoke a sense of complexity and interdependencies (but generally only a small piece at a time)
- Offers interactive exploration through a semi-immersive visualization

## SOME VARIATIONS

- **Sentiment Analysis:** Using Concept Lists / Union Concept Lists (generalization thesauruses) to categorize sentiment based on common phrases in the text corpuses -> general sentiments
- **The State of a Field:** Application of network-based text analysis to the entire corpuses of domain fields to understand salient concepts (or clusters of studies) in that field ...and analyses of bibliographies and references lists (for co-author networks and related topics and years of varying productivity)

# A REVIEW OF THE PROCESS



AutoMap Autocoding and ORA NetScenes Visualization for Network-Based Text Analysis  
(One Conceptualization of the Process)

# A REVIEW OF THE PROCESS

- Objectives
- Research Design
- Institutional Review Board (IRB) Process
- Text Corpus Selection
- Stopwords or Delete Lists
- Concept Lists / Union Concept Lists (Generalization Thesauruses)
  - Machine-generated ones tend to be quite inclusive and complex (requiring whittling down)
- Text Processing
- Graph Visualizations Created
- Human-Based Analysis

# POTENTIALS?

- What are some ways that these software tools and methods may be used in your areas of expertise?
- Any hesitations? Concerns?



The page features decorative circuit-like lines in the corners. The top-left and top-right corners have black lines, while the bottom-left and bottom-right corners have gold lines. The background is a light gray with a pattern of concentric circles.

# FOR RESEARCHERS

AUTOMAP / ORA-NETSCENES



# CURRENT RESEARCH

## **Textually**

- Applications to computational folkloristics and domain-specific knowledge structures
- Applications to sentiment analysis from keyword extractions from microblogging corpuses (conversations, account Tweet streams) and Web networks

## **Organizationally**

- Applications to terror networks, covert networks, and law enforcement interests
- Applications to community resilience for emergency responses




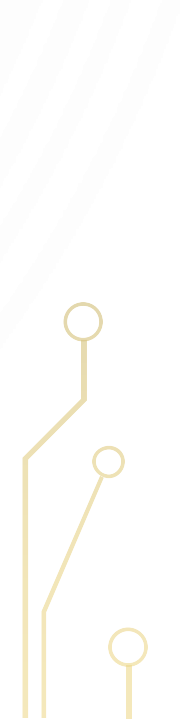


# LEARNING THE METHODS AND TOOLS

- Understand the rationale for the tools. (There are plenty of widely available publications related to the tools' origins and the rationales of the development teams.)
- Acclimate to the tools first and actually sample their capabilities in depth (by using a simple text corpus initially).
- It's critical to know the datasets well. This includes knowledge of where the data came from, how it's cleaned, and other aspects.



## LEARNING THE METHODS AND TOOLS (CONT.)

- Automation of text mining provides different insights than manual reading. Such machine-based text analysis offers some types of knowledge and not others. When making assertions, it's important to be nuanced and clear about what is being asserted (gist, broad themes, possible relationships between semantics and concepts) and what isn't (in-depth and fine-tuned insights).
  - Text network analysis offers evocative visualizations in both 2D and 3D. Visualizations are eye-catching, but they have to be presented properly to be valuable in an information setting.
- 
- 

## LEARNING THE METHODS AND TOOLS (CONT.)

- There may be machine-based artifacts in terms of the data. This is why it's critical to understand both the text corpuses and the software tools thoroughly. It is important to explain the findings with clarity (to avoid misconceptions).
- Both software tools can ingest a wide range of data. AutoMap can ingest text data; it can ingest all sorts of thesauruses. ORA NetScenes can ingest various DyNetML files to output graph visualizations.



## LEARNING THE METHODS AND TOOLS (CONT.)

- Researchers will need to bring their domain knowledge to bear on the tool and the network-based text analysis findings. Generally, this does not work as a stand-alone tool but a complementary one with other research methods and insights.
- It helps to read up on the research literature to get a sense of the language used around assertions of analytical value based on the graphs and graph metrics.
- It helps to read up on dynamic network-based text analysis to see how this is applied in even more challenging real-time contexts.

# DESIRABLE RESEARCHER SKILLSETS

- Grammar and syntax (linguist skills)
- Mid-level statistical savvy
- Mid-level computational skills and methodical step-by-step approaches
- Basic data structures and handling of datasets
- Graph analysis and spatial reasoning (the identification of patterns)
- Wide reading
- Traits: Patience and doggedness (even if forced by self-will)

The page features decorative circuit-like lines in the corners. The top-left and top-right corners have black lines, while the bottom-left and bottom-right corners have gold lines. All lines end in small circles, resembling nodes or components on a circuit board. The background is a light gray with a subtle pattern of concentric circles.

# SOME RESOURCES

# RELATED LINKS

## Software Downloads

- AutoMap: <http://www.casos.cs.cmu.edu/projects/automap/>
- ORA Software: <http://www.casos.cs.cmu.edu/projects/ora/software.php>

## Support Group

- ORA Google Group:  
<http://www.casos.cs.cmu.edu/projects/automap/ORAGoogleGroup.php>

## Originating Organization

- Center for Computational Analysis of Social and Organizational Systems (CASOS of Carnegie Mellon University): <http://www.casos.cs.cmu.edu/>

# OTHER GRAPH VISUALIZATION TOOLS

## Free and Open-Source Tools

- NodeXL (Network Overview, Discovery and Exploration for Excel): <http://nodexl.codeplex.com/> (of CodePlex)
- Pajek: <http://pajek.imfm.si/doku.php>
- Gephi: <https://gephi.org/>

## Commercial Tool

- UCINET and NetDraw (Analytic Technologies): <http://www.analytictech.com/>





## REFERENCES

- Bigrigg, M.W., Carley, K.M., Kunkel, F., Diesner, J., Eisenberg, T., Chieffallo, D., & Columbus, D. (2010). AutoMap: Extracting usable information from unstructured texts.
- Sangal, A., Carley, K.M., Altman, N., & Martin, M.K. (2012). Creating, using and updating thesauri files for AutoMap and ORA. CASOS Technical Report: CMU-ISR-12-108.

# CONCLUSION AND CONTACT

- Dr. Shalin Hai-Jew
    - Instructional Designer
  - Information Technology Assistance Center
  - Kansas State University
  - 212 Hale Library
  - 785-532-5262
  - [shalin@k-state.edu](mailto:shalin@k-state.edu)
- 
- The presenter has no ties with CASOS or Carnegie Mellon University.

