Creating, Using and Updating Thesauri Files For AutoMap and ORA

Abhinav Sangal, Kathleen M. Carley,
Neal Altman, Michael K. Martin

July 26, 2012
CMU-ISR-12-108

Institute for Software Research
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Contact Information:
{sangala}@andrew.cmu.edu,
{carley, na22, mkmartin}@cs.cmu.edu

Center for the Computational Analysis of Social and Organizational Systems

CASOS technical report.

This work was supported in part by the Office of Naval Research – ONR -N000140811223. (SORASCS), ONR -000140910667 (CATNET), ONR - N000140811186 (Ethnographic). Additional support was provided by the center for Computational Analysis of Social and Organizational Systems (CASOS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the Department of Defense or the U.S. government.

Keywords: Automap, ORA, Thesaurus, Universal Thesaurus, Meta-ontology, Domain Thesaurus, Split, Merge, Algorithm.

Abstract

AutoMap [1] is text analysis software that performs Network Text Analysis by running an automated process on a corpus of raw text data to generate one or more meta-networks which include the nodes and links representing relations among entities described. Automap uses thesaurus files [1] when creating meta-networks. These thesaurus files are list which allows the association of words or phrases found in texts with abstract concepts and/or node classes used in the extracted meta-networks.

Over time, a large number of thesauri have been created. Many of the extant thesauri contain entries that are relevant to new text analysis projects. But thesaurus re-use is difficult due to the number of thesauri. In this report, we describe one approach to making thesaurus re-use easier by combining and reconciling multiple thesauri into one under user control.

With this approach, the process of creating a Meta network out of a raw corpus of text data is more efficient and the user is able to perform a more accurate analysis of the Meta network, as the individual thesauri files can be merged to create a single and large Universal or Master Thesaurus containing all the general abstract concepts, along with several different Domain-specific thesauri.

In the following report, we first discuss the differences between a Universal thesaurus and the domain or the project specific thesauri. We then go on to discuss the evolution in the formats of the thesauri used by AutoMap, followed by a discussion of the standard Dynamic Network Analysis (DNA) meta-ontology [1].

We then detail the process used to create a single universal/master thesaurus and several different Domain thesauri. The process involves a mix of two major processes which we refer to as the Split routine and the Merge routine. We shall discuss the Split routine and the merge routine algorithm along with the process that has been used to merge and create a single thesaurus file by combining a large number of thesauri files. The merge process is not a simple process of combining all the files into one file; it involves some computational functions to make this process more efficient and more accurate. These functions are deleting duplicates, detecting the concept cycles and performing a depth first search for each concept.

The paper concludes by discussing some future improvements which could be made to the process so as to improve and automate the process which is being used at present for the merge and split process.

Table of Contents

1 Introduction

AutoMap [1] is software tool for computer-assisted Network Text Analysis (NTA). NTA encodes the links among concepts in a text and constructs a network of the links among concepts. AutoMap subsumes classical Content Analysis by analyzing the existence, frequencies, and covariance of terms and themes.

For the purpose of NTA and in order to generate a Meta network from a corpus of raw data, AutoMap uses some files for reference. The files are referred to as the Thesaurus files. Thesaurus files are essentially lists of words in comma separated values (.csv) format. The thesaurus files are used for many purposes during the text analysis process. They can be used to create a delete list, can contain a list of noise words for filtering and even project specific concepts.

Thesauri files are an integral part of the network analysis procedure. Over time, many thesauri have been developed. As more thesauri are created, managing them has become progressively more difficult. To improve and polish the process of network analysis, we need to create a better and more efficient thesaurus.

In order to create a more efficient thesaurus, one approach is to merge all the existing files. But this is not sufficient. We have to merge the files in a way that no entry is duplicated and no concept cycles are formed. Concept cycles are the relational cycles which are formed while performing the merge operation on the thesauri files. For example: A maps to B and then B maps to C and then we find that C maps to A, so we say that this is a concept cycle and we just map A to A in the thesaurus then. For specific bodies of text, we thought that it would be useful to create supplemental domain thesauri files separate from the universal thesaurus because these domain thesauri contain entries that are not universally relevant. The major difference between the Universal thesaurus and the domain thesauri is that the universal thesauri contains general abstract concepts which may be useful in almost every project relating to generate a network from the data set. In contrast, the domain thesauri can be visualized as the project specific thesauri which contain concepts which are specific to the project. Hence, to enhance the efficiency of creating the Meta network, we incorporate a Split routine along with the merge process.

The approach for differentiating the universal thesaurus from the domain thesaurus can vary from person to person. In this report we differentiate the domain and the universal thesaurus by distinguishing the concepts which are single word agents from the others. For example, consider a data set on American politics. Now the concept entry, **Obama** shall go to the domain thesauri since this concept is a one word agent without any white space. Whereas, the concept **Barack Obama** should go the universal thesaurus since it is a two-word agent containing white spaces

The following section starts by discussing the functions of thesauri and the types into which it can be classified according to the varying file formats, functions and according to the type of concept entries it contains. We discuss the types of thesauri based on the format and how the evolution took place with regard to the formats of the thesauri files. We continue by explaining what each of the columns contain and their meaning. Each concept in the data set can be classified into a semantic category (i.e., a node class representing a particular type of entity) according to the context in which it is used in the texts; this could be agent, organization, location, resource, event etc. The aggregation of these categories is known as the meta ontology. We also briefly discuss the various meta ontologies and how each concept can be classified into the correct ontology or can be deleted and ignored.

After discussing the format and purpose of the thesaurus, we discuss the difference between the universal thesaurus and the domain thesauri. We also discuss the precedence of the domain thesauri over the universal thesaurus and how this can be used in AutoMap during meta-network extraction. Next, we discuss the split and the merge routines and the simplified process used for carrying out the merge and split process. Also, we discuss the algorithm for the split routine and the merge routine in a very simple computing linguistics.

Finally, we conclude the report by discussing the benefits and some problems which the present split and merge process provides. Also, we discuss the various other future improvements which should be made to this process in order to achieve an efficient universal thesaurus and hence a relevant Meta network from the data set.

2 AutoMap Thesauri

A thesaurus is a list which associates actions or abstract concepts with words or phrases found in text. Apart from this, a thesaurus can be used to provide additional information about the listed concepts. A thesaurus can be created by AutoMap or manually with a text editor as a text file (.txt) or a spreadsheet program that can save as (.csv) files.

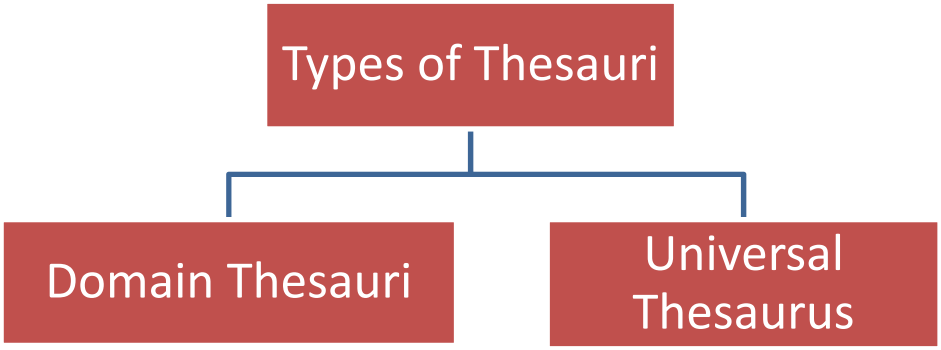


Figure 1 Major types of thesauri.

3 Thesauri Format Types

Thesauri can be classified into six types on the basis of the format of the thesaurus, these include the single column format, 2 column generalization format, 2 column meta network format, reduced format, change or review format. A thesaurus file can be in one of six types:

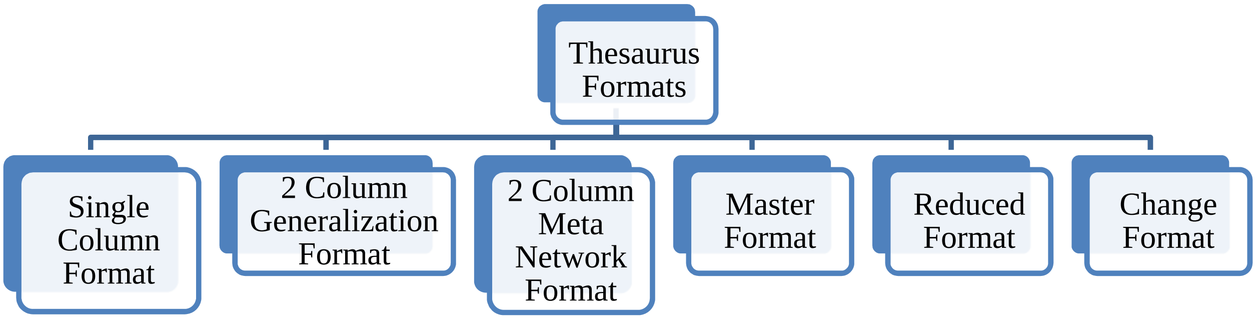
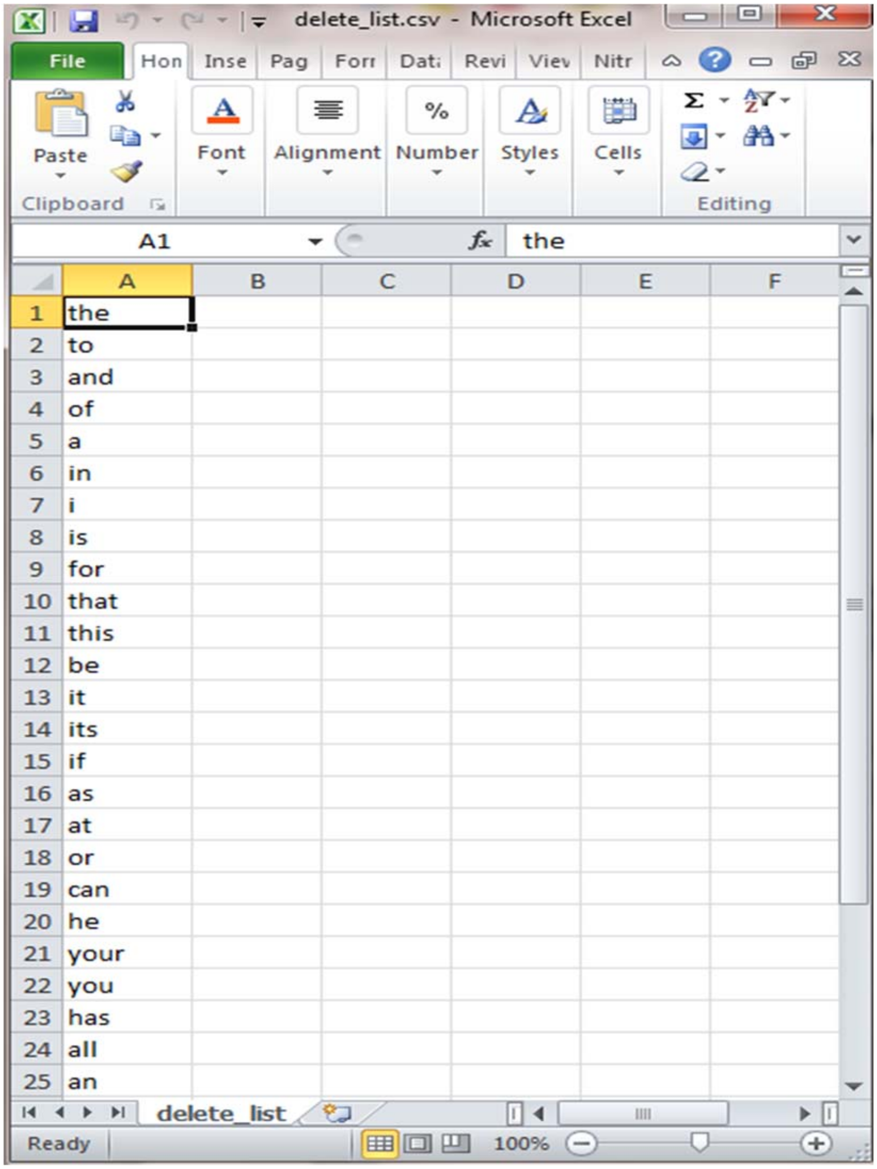


Fig 2: Thesaurus formats

3.1 Single Column Format

The single column format is the least complex of the thesaurus formats. It consists of only one column, without a column header. The single column format is primarily used for delete lists. Delete lists are lists which when used as input thesauri to AutoMap, delete any of the concepts listed from the data set so they will not be included while generating the Meta network.

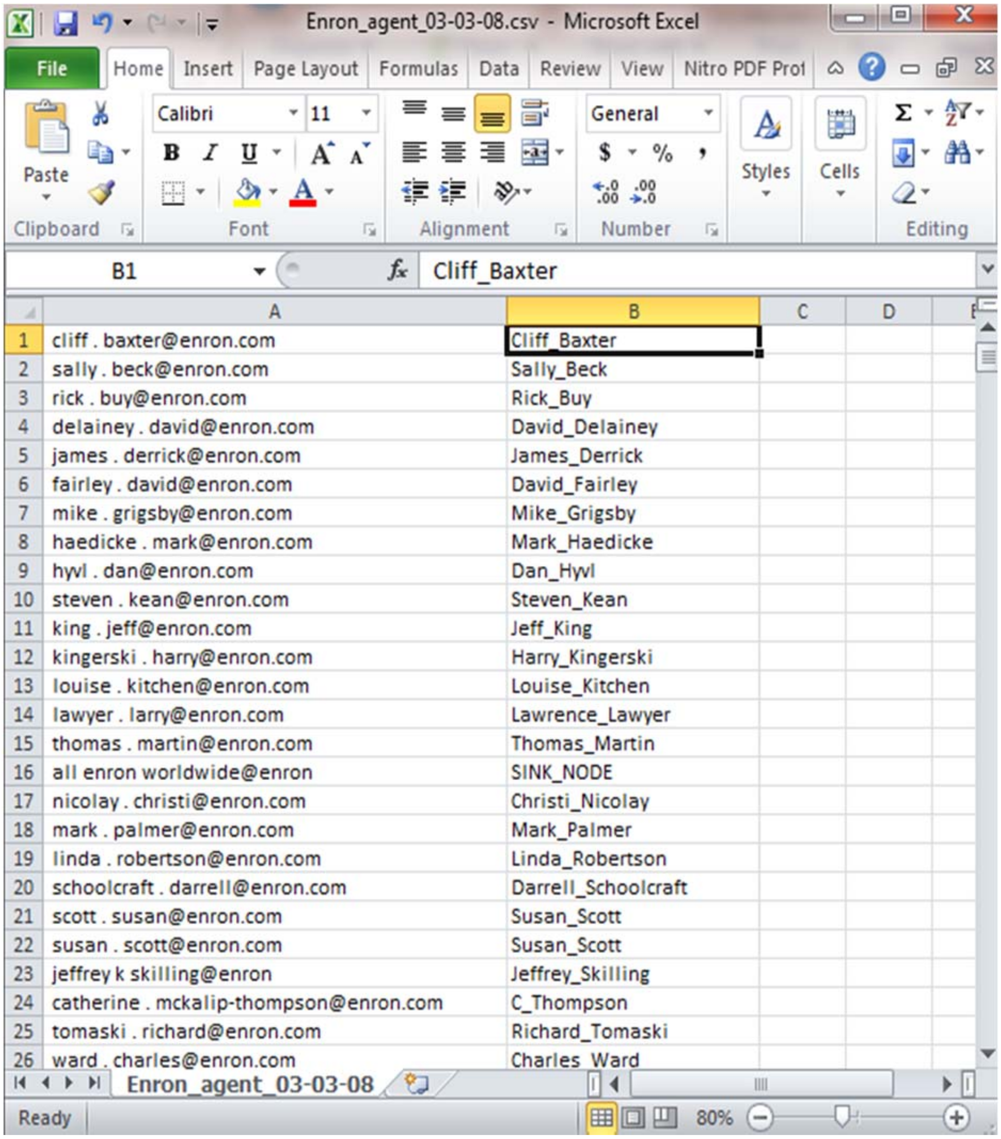


	A	B	C	D	E	F
1	the					
2	to					
3	and					
4	of					
5	a					
6	in					
7	i					
8	is					
9	for					
10	that					
11	this					
12	be					
13	it					
14	its					
15	if					
16	as					
17	at					
18	or					
19	can					
20	he					
21	your					
22	you					
23	has					
24	all					
25	an					

Fig 3: Single column format type.

3.2 Two Column Generalization Format

Typically a two-column thesaurus is used to associate text-level concepts with higher-level concepts, where the text-level concepts represents the content of a data set, and the higher level concepts represent the text-level concepts in a generalized way. No column headers are included.



Cliff_Baxter

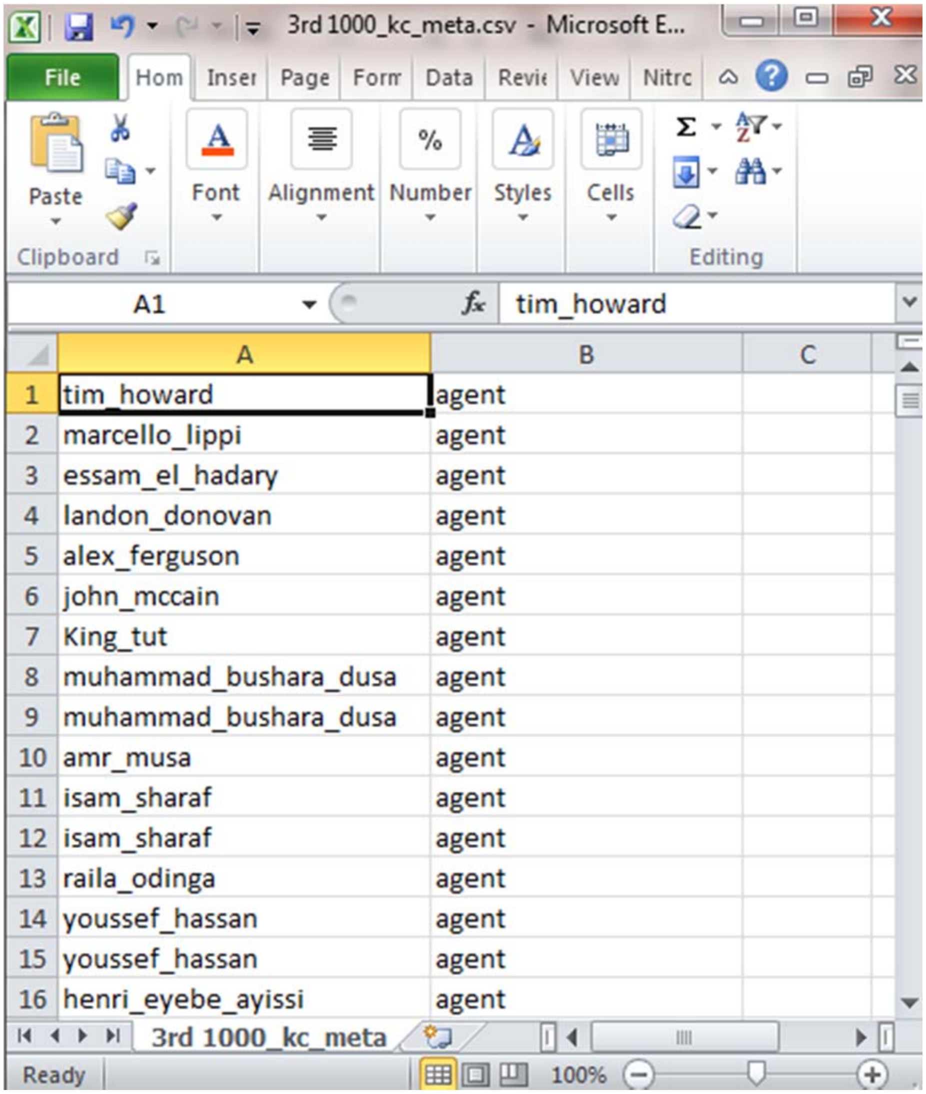
	A	B
1	cliff . baxter@enron.com	Cliff_Baxter
2	sally . beck@enron.com	Sally_Beck
3	rick . buy@enron.com	Rick_Buy
4	delainey . david@enron.com	David_Delainey
5	james . derrick@enron.com	James_Derrick
6	fairley . david@enron.com	David_Fairley
7	mike . grigsby@enron.com	Mike_Grigsby
8	haedicke . mark@enron.com	Mark_Haedicke
9	hyvl . dan@enron.com	Dan_Hyvl
10	steven . kean@enron.com	Steven_Kean
11	king . jeff@enron.com	Jeff_King
12	kingerski . harry@enron.com	Harry_Kingerski
13	louise . kitchen@enron.com	Louise_Kitchen
14	lawyer . larry@enron.com	Lawrence_Lawyer
15	thomas . martin@enron.com	Thomas_Martin
16	all enron worldwide@enron	SINK_NODE
17	nicolay . christi@enron.com	Christi_Nicolay
18	mark . palmer@enron.com	Mark_Palmer
19	linda . robertson@enron.com	Linda_Robertson
20	schoolcraft . darrell@enron.com	Darrell_Schoolcraft
21	scott . susan@enron.com	Susan_Scott
22	susan . scott@enron.com	Susan_Scott
23	jeffrey k skilling@enron	Jeffrey_Skilling
24	catherine . mckalip-thompson@enron.com	C_Thompson
25	tomaski . richard@enron.com	Richard_Tomaski
26	ward . charles@enron.com	Charles_Ward

Fig 4: The two column generalization format type.

3.3 Two Column Meta Network Format

This is a variation of the basic two column thesaurus which has the concept in one column and the meta ontology (see section 5 Meta Ontologies) in the other one. Associates text level concepts with the Meta network category (i.e., node class).

This differs from the two column generalization thesaurus in the sense that the second column is the meta ontology rather than the conceptTo column in generalization thesaurus. Note however the regular two columns and meta network format can only be determined by inspection of the second column or by file naming conventions.

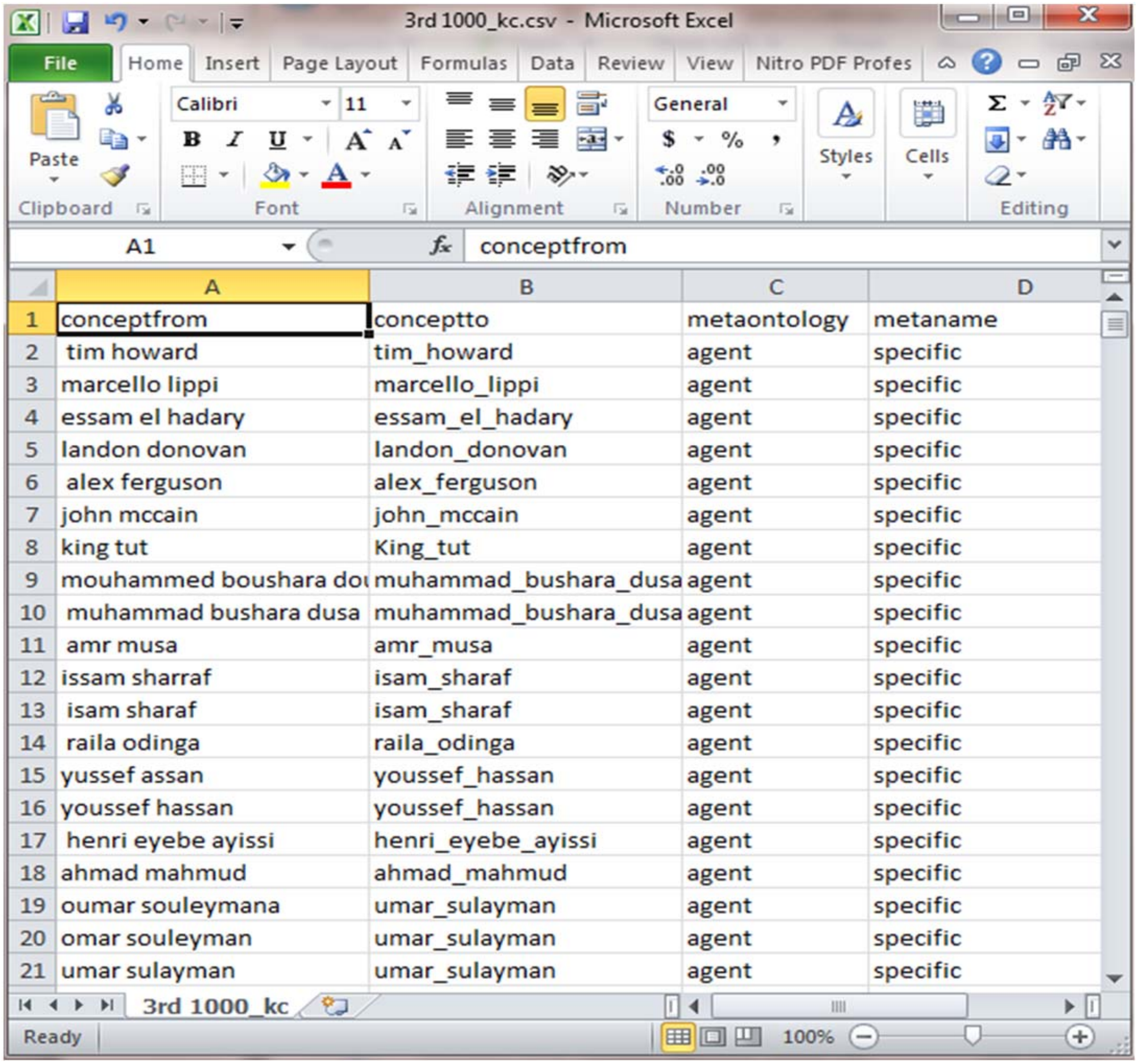


	A	B	C
1	tim_howard	agent	
2	marcello_lippi	agent	
3	essam_el_hadary	agent	
4	landon_donovan	agent	
5	alex_ferguson	agent	
6	john_mccain	agent	
7	King_tut	agent	
8	muhammad_bushara_dusa	agent	
9	muhammad_bushara_dusa	agent	
10	amr_musa	agent	
11	isam_sharaf	agent	
12	isam_sharaf	agent	
13	raila_odinga	agent	
14	youssef_hassan	agent	
15	youssef_hassan	agent	
16	henri_eyebe_ayissi	agent	

Fig 5: Two column Meta network format type.

3.4 Master Format

The master format contains the same information as the two-column format plus additional attribute information. It associates the text-level concepts with higher-level concepts, where the text-level concepts refer to the content of a data-set and the higher level concepts represent the text-level concepts in a generalized way. In addition the, master thesauri format contains metaOntology column which describes the node class to which a particular concept can be categorized (for example, Barack Obama maps to agent and India maps to location). The metaName column defines whether the concept from the text is a generic concept or a specific one (see section 6 Generic and specific). The master thesaurus format thus combines the two column formats with additional information.



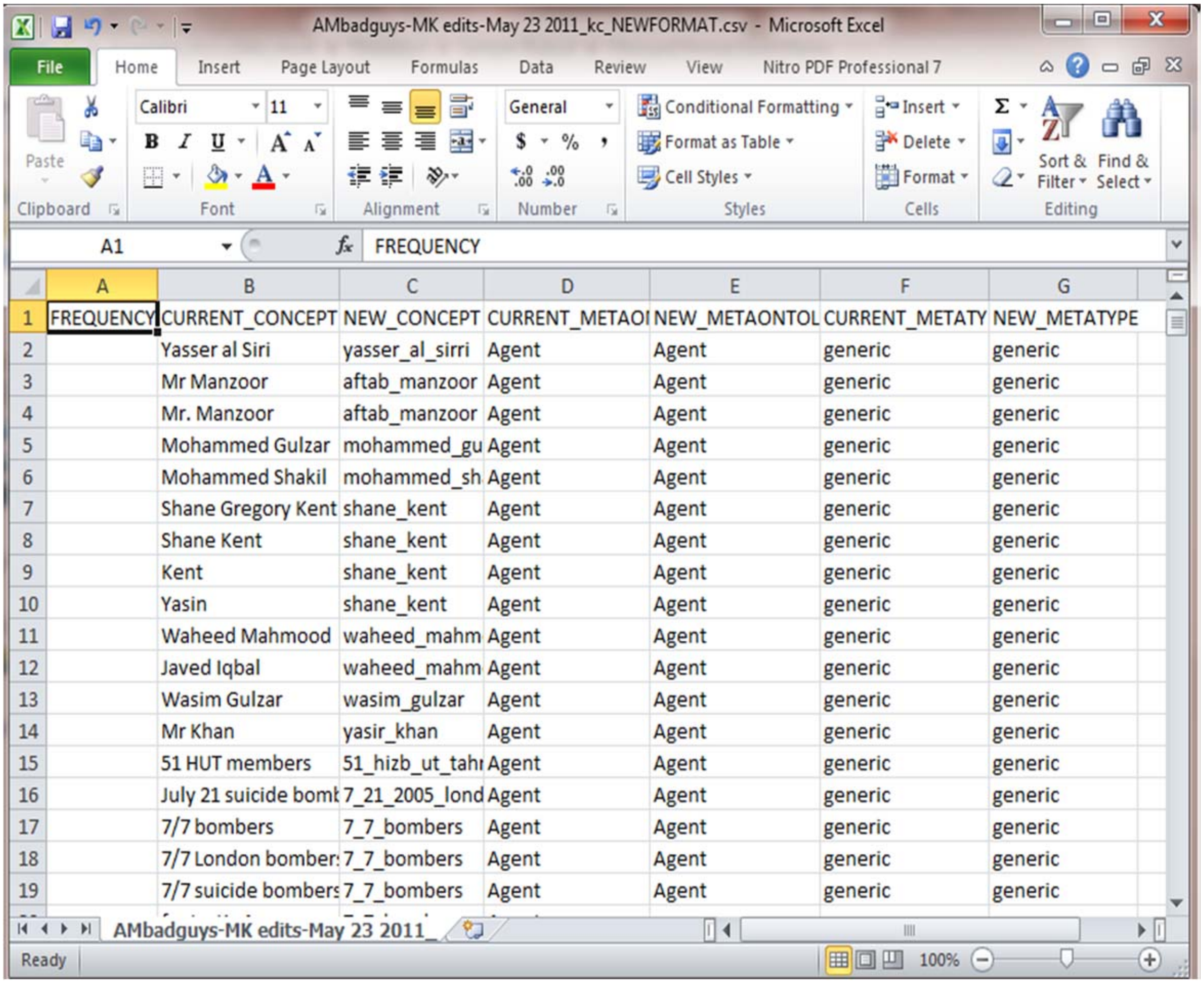
A1 *fx* conceptfrom

	A	B	C	D
1	conceptfrom	conceptto	metaontology	metaname
2	tim howard	tim_howard	agent	specific
3	marcello lippi	marcello_lippi	agent	specific
4	essam el hadary	essam_el_hadary	agent	specific
5	landon donovan	landon_donovan	agent	specific
6	alex ferguson	alex_ferguson	agent	specific
7	john mccain	john_mccain	agent	specific
8	king tut	King_tut	agent	specific
9	mouhammed boushara dou	muhammad_bushara_dusa	agent	specific
10	muhammad bushara dusa	muhammad_bushara_dusa	agent	specific
11	amr musa	amr_musa	agent	specific
12	issam sharraf	isam_sharaf	agent	specific
13	isam sharaf	isam_sharaf	agent	specific
14	raila odinga	raila_odinga	agent	specific
15	yussef assan	youssef_hassan	agent	specific
16	youssef hassan	youssef_hassan	agent	specific
17	henri eyebe ayissi	henri_eyebe_ayissi	agent	specific
18	ahmad mahmud	ahmad_mahmud	agent	specific
19	oumar souleymana	umar_sulayman	agent	specific
20	omar souleyman	umar_sulayman	agent	specific
21	umar sulayman	umar_sulayman	agent	specific

Fig 6: Master thesauri format.

3.5 Reduced Format

This format subsumes the previous 3 formats with respect to the data it contains. There are 7 columns: Frequency, cuurent_concept, new_concept, current_metaontology, new_metaontology, current_metatype, new_metatype. This format provides users with a means to change meta-network structure (i.e., nodes and/or links) by deleting, generalizing, or re-classifying nodes (i.e., entities extracted from text). The meta-type attribute value may also be changed. For example: If a concept USA has been classified as location and the user wants this concept to be an Organization instead, the user can change the meta-ontology by putting the new meta ontology in the **new_metaontology** column (see sections 5 Meta Ontologies on page 12 and 6 Generic and specific on page 14). This format is also known as the Review format.



A1 FREQUENCY

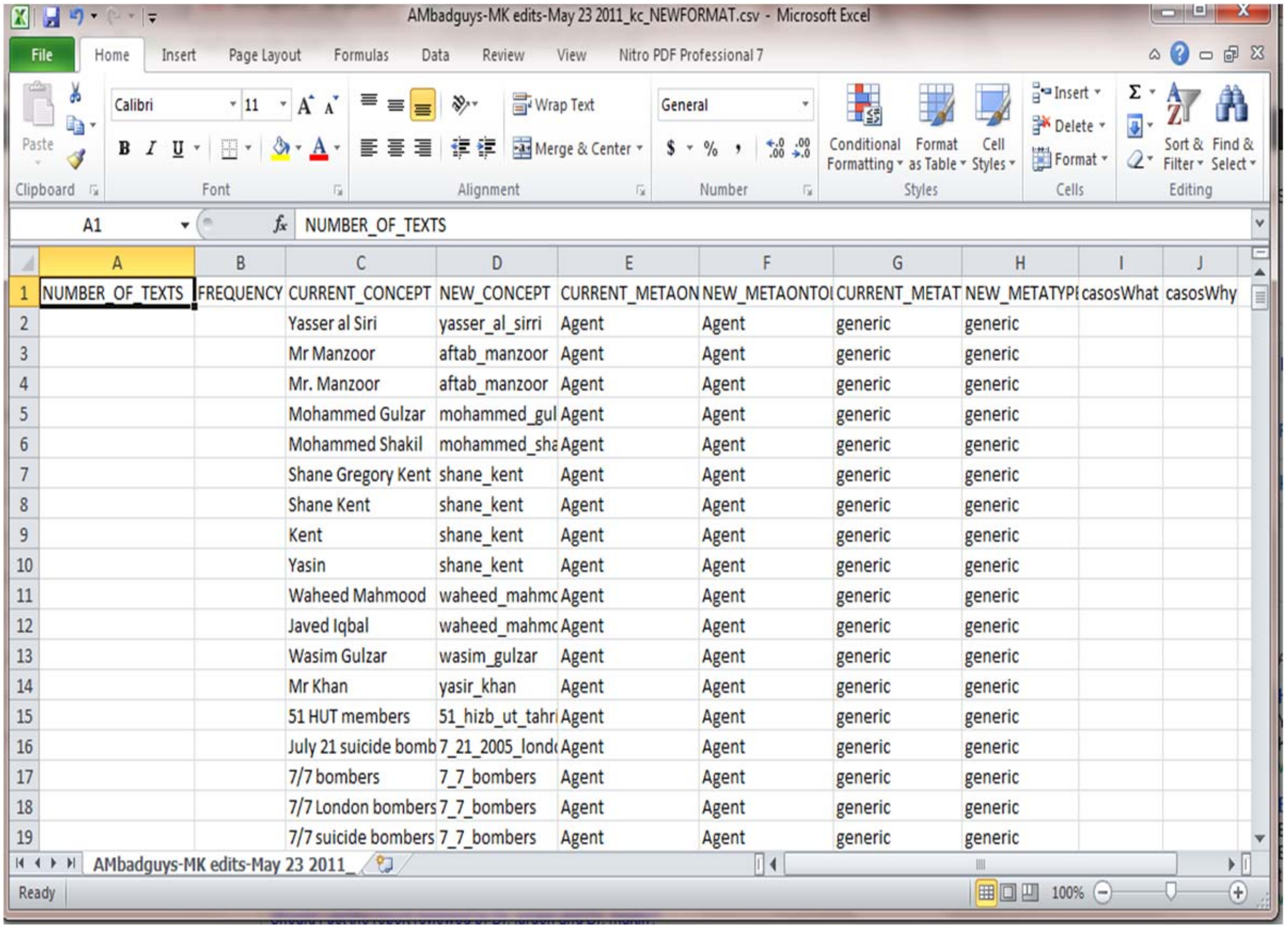
	A	B	C	D	E	F	G
1	FREQUENCY	CURRENT_CONCEPT	NEW_CONCEPT	CURRENT_METAOI	NEW_METAONTOL	CURRENT_METATY	NEW_METATYPE
2		Yasser al Siri	yasser_al_sirri	Agent	Agent	generic	generic
3		Mr Manzoor	aftab_manzoor	Agent	Agent	generic	generic
4		Mr. Manzoor	aftab_manzoor	Agent	Agent	generic	generic
5		Mohammed Gulzar	mohammed_gu	Agent	Agent	generic	generic
6		Mohammed Shakil	mohammed_sh	Agent	Agent	generic	generic
7		Shane Gregory Kent	shane_kent	Agent	Agent	generic	generic
8		Shane Kent	shane_kent	Agent	Agent	generic	generic
9		Kent	shane_kent	Agent	Agent	generic	generic
10		Yasin	shane_kent	Agent	Agent	generic	generic
11		Waheed Mahmood	waheed_mahm	Agent	Agent	generic	generic
12		Javed Iqbal	waheed_mahm	Agent	Agent	generic	generic
13		Wasim Gulzar	wasim_gulzar	Agent	Agent	generic	generic
14		Mr Khan	yasir_khan	Agent	Agent	generic	generic
15		51 HUT members	51_hizb_ut_tahı	Agent	Agent	generic	generic
16		July 21 suicide bomb	7_21_2005_lond	Agent	Agent	generic	generic
17		7/7 bombers	7_7_bombers	Agent	Agent	generic	generic
18		7/7 London bomber:	7_7_bombers	Agent	Agent	generic	generic
19		7/7 suicide bombers	7_7_bombers	Agent	Agent	generic	generic

Fig 7: Reduced format thesauri.

3.6 Change Format

Comprises of 10 columns and subsumes all the previous formats of Thesauri. Out of these 10 columns, seven columns are same as that of the reduced format. The three other columns are: casosWhat, casosWhy, and number_of_texts. The casosWhat column contains the name of the thesaurus file from where the concept comes. The casosWhy column essentially is an explanation for the mapping from the Initial concept to the New concept. For example: “Entry created because of depth first search”. The number_of_texts column, as the name suggests contains a numeric value relating to how many texts contain the concept.

Note: There can be other columns also, apart from the ones that are listed. For example: TF-IDF, EPA etc.



fx NUMBER_OF_TEXTS

	A	B	C	D	E	F	G	H	I	J
1	NUMBER_OF_TEXTS	FREQUENCY	CURRENT_CONCEPT	NEW_CONCEPT	CURRENT_METAON	NEW_METAONTOI	CURRENT_METAT	NEW_METATYPI	casosWhat	casosWhy
2			Yasser al Siri	yasser_al_sirri	Agent	Agent	generic	generic		
3			Mr Manzoor	aftab_manzoor	Agent	Agent	generic	generic		
4			Mr. Manzoor	aftab_manzoor	Agent	Agent	generic	generic		
5			Mohammed Gulzar	mohammed_gul	Agent	Agent	generic	generic		
6			Mohammed Shakil	mohammed_sha	Agent	Agent	generic	generic		
7			Shane Gregory Kent	shane_kent	Agent	Agent	generic	generic		
8			Shane Kent	shane_kent	Agent	Agent	generic	generic		
9			Kent	shane_kent	Agent	Agent	generic	generic		
10			Yasin	shane_kent	Agent	Agent	generic	generic		
11			Waheed Mahmood	waheed_mahmc	Agent	Agent	generic	generic		
12			Javed Iqbal	waheed_mahmc	Agent	Agent	generic	generic		
13			Wasim Gulzar	wasim_gulzar	Agent	Agent	generic	generic		
14			Mr Khan	yasir_khan	Agent	Agent	generic	generic		
15			51 HUT members	51_hizb_ut_tahri	Agent	Agent	generic	generic		
16			July 21 suicide bomb	7_21_2005_lond	Agent	Agent	generic	generic		
17			7/7 bombers	7_7_bombers	Agent	Agent	generic	generic		
18			7/7 London bombers	7_7_bombers	Agent	Agent	generic	generic		
19			7/7 suicide bombers	7_7_bombers	Agent	Agent	generic	generic		

Fig 8: Changed format thesauri.

3 Evolution of the AutoMap Thesaurus

Over a period of years and as part of the process of making network analysis more detailed and informative, the format of the thesaurus has evolved according to the needs of the user.

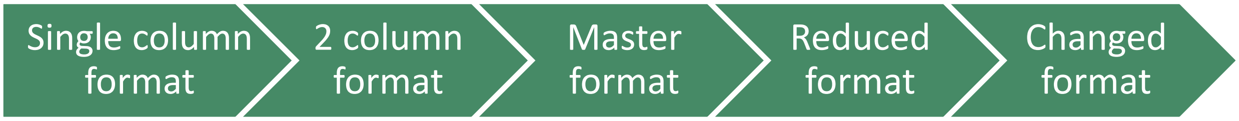


Fig 9: Evolution of Thesaurus

1. **Single column format**: One of the earliest types of the thesaurus file used in the Automap for network analysis. File using this format remain in use for delete lists in AutoMap. It supports deleting unneeded concepts from the raw data.

2. **Two column format**: The two column format was an elaboration over the single column type format. This format is usually used for two types of thesauri, namely, 2 column generalization formats and 2 column meta network format. The 2 column generalization format basically relates the concepts from the data set to more abstract concepts. Whereas, the 2 column meta network thesaurus relates the concepts in the data to their ontology.

3. **Master format:** Master format combines the two types of the 2 column formats and contains 4 columns and adds meta name (or meta type) information. The four columns are conceptFrom, conceptTo, metaOntology, metaName. The master format can be used as a delete list and a meta network thesaurus as well.

4. **Reduced format:** The reduced format was a more advanced type of format and was basically developed for the purpose of performing editorial and deletion changes easily using ORA [2]. For this purpose, 2 new columns were added to the master format, which were the new metaOntology and the new metaType in addition to the previous ones. Apart from using reduced list as a simple delete list and a meta network thesaurus, it can be used directly in ORA to modify the links and the nodes using the #delete and #merge actions.

5. **Change format:** At present, the most advanced and subsumes all the previous formats. The changed format is the 10 column format which includes 3 extra columns apart from the reduced format type. These additional columns are casosWhat, casosWhy and

6. **NUMBER_OF_TEXTS** columns. This format like the reduced format can be used for modifying the links and the nodes directly in ORA [2].

4 Thesaurus Columns Defined

In this section, we discuss what each column means in the changed format:

1. **NUMBER_OF_TEXTS:** This column contains a numerical value which relates to how many times each concept occurs in a corpus of text.For example: If the concept **Honda** comes in 5 text sets, this column would contain a number

2. **FREQUENCY:** This column contains a numerical value pointing to how frequent the concept is in a corpus of text data.

3. **CURRENT_CONCEPT:** The set of words you want converted/ translated into NEW_CONCEPT. The CURRENT_CONCEPT can be a phrase in the article you are coding or a concept in dynetml.

4. **NEW_CONCEPT:** These are the words/ n-grams which are mapped from the CURRENT_CONCEPT column, these entries are the used as the final concepts in the Meta network. The blanks are converted into underscores in the new concept column. No capitalization is present in this column, and there should be no symbols.

5. **CURRENT_METAONTOLOGY:** This column is generated by the . AutoMap classifies the concept term into one of the Meta Ontologies values or leaves it blank. The allowable meta-ontology terms are: agent, organization, location, task, event, knowledge, resource, belief.

6. **NEW_METAONTOLOGY:** If the ontology class deduced by AutoMap is wrong then put the right class value in the column NEW_METAONTOLOGY. If the concept under CURRENT_CONCEPT should be deleted and not coded then under NEW_METAONTOLOGY put #delete. If the phrase under CURRENT_CONCEPT should not be treated as a phrase but split into its parts then under NEW_METAONTOLOGY #split.

7. **CURRENT_METATYPE:** This column specifies the type for CURRENT_CONCEPT as defined by AutoMap. It is either "generic" or "specific" or blank. Note **ONLY** agent, organization, location, event can have a type. The various metaOntology listed above can be classified into generic and specific type. For example: **John Doe** is a specific agent whereas **policeman** is a generic agent.

8. **NEW_METATYPE:** If the type deduced by the AutoMap seems wrong or if there is no type and the concept is an agent, organization, location, event then the user can add specify the type under NEW_METATYPE.

5 Meta Ontologies

The Ontology is defined as the group of classes in which almost every concept and phrase can be classified depending on the context in which the concept/phrase is used in a particular sentence or a paragraph. For example: USA can be a location or an Organization, depending on the context of the data.

5.1 Standard Node Classes

In detail, the ten entity classes can be defined as follows:

1. **Agents** are individual decision makers. The most common type of decision makers is people. However, this category could also be used for other types of actors such as robots or monkeys. A single *Agent* represents any person: a family member, a Roman soldier, Soothsayer, Calpurnia, Cicero, any historical figure, a terrorist, or a teacher.

2. **Resources** are products, materials, or goods that are necessary to perform *Tasks*, *Events*, and *Actions*. A Resource could literally be a dagger, a cloak, a crown, a short sword, a computer, money, bombs, tools, or books.

3. **Knowledge** describes cognitive capabilities and skills. *Knowledge* could be Trigonometry, History, *English*, Economics, the science of DNA, or the knowledge about how to perform a surgery or to build bombs.

4. **Tasks** are well defined procedures or goals. A *Task* could be any process in a company, e.g. product development or administration, but also the plan to kill Caesar.

5. **Organizations** are collectives of people that try to reach a common goal. An Organization could be a specific company, the United Nations, or the government of a country.

6. **Locations** are geographical positions at any aggregation level that describe places or areas. A *Location* could be 1600 Pennsylvania Avenue in Washington, London, Florida, The Middle East, Earth, or Mars.

7. **Events** are occurrences or phenomena that happen. An *Event* could be 9-11, the JFK Assignation, the Super Bowl, a wedding, a funeral, or an inauguration. Specific events are one time occurrences with a specific date.

8. **Actions** are specific activities done by agents. An *Action* could be buying a car, writing a letter of recommendation, or flying to Africa.

9. **Beliefs** are any form of religion or other persuasion. A Belief could be to believe that there is a god, or that there are many gods, or that that Earth is flat. Some beliefs are signaled by sentiment such as "war is bad."

10. **Roles** describe functions of individual decision makers abstracted from specific *Agents*. A *Role* could be leader of a group, driver of a car, or mother of an *Agent*.

11. **Groups** also referred to as meta-nodes, are any categorization of nodes into a cluster. Groups are defined by one or more of the grouping algorithms.

To repeat the function of these node classes and to show their differences as well as give a first impression about how these node classes are connected with each other, we summarize the Julius Caesar plot very briefly. Brutus (*Agent*) and other senators (*Roles*) agree that Rome (*Location*) would be a better place without Caesar (*Belief*). To kill Caesar (*Task*) they form a group of assassins (*Organization*). To accomplish their task they need to know about Caesar's daily routine (*Knowledge*) and how to get their knives (*Resources*) into the senate. Finally, the assassination (*Event*) takes place because somebody actually stabs Caesar (*Action*).

5.2 Actions

Apart from the various meta ontologies, the NEW_METAONTOLOGY column in the reduced format and the change format can contain some actions which help in rearranging the links and managing the nodes directly in ORA [2]. These actions help the users to manually edit the thesaurus according to their own needs.

1. **Delete:** This is actually an action rather than an ontology item, but this is placed in the Meta ontology column for the convenience of the user. The user can delete an entry by writing #delete in the New Meta ontology column if the user wishes to delete the entry. When the #delete tag is added to the column, that concept is deleted from the text.

2. **Ignore:** This is also an action rather than a Meta Ontology value. Ignore is used if the set of concepts should be treated as individual words. The user can use this facility by writing #ignore in front of the entry which has to be ignored. When the #ignore tag is added to the metaOntology column, that particular concept is ignored.

3. **Split:** This is also an action similar in usage to the Delete and Ignore actions. If a group of words are mapped as a phrase, and if the user wants the words in phrase to be independent concepts, then the user can add #split tag in the New_metaOntology column.

6 Generic and specific

The generic and specific are the two allowed **metaType** values**.** They provide the users with some additional information on whether this is a universal or general concept or a specific concept. The difference between Specific and Generic can be simply put as the difference between **a** and **the.**

1. **Specific:** Concepts which reference a specific agent, location or organization. For example: **John Doe** is a specific agent, **CASOS** is a specific organization, and **USA** is a specific location.

2. **Generic**: concepts which are universal and do not point to a specific agent, location or an organization. For example: **man** is a generic agent, **health fund organization** is a generic organization, and **my room** is a generic location.

7 Delete Lists

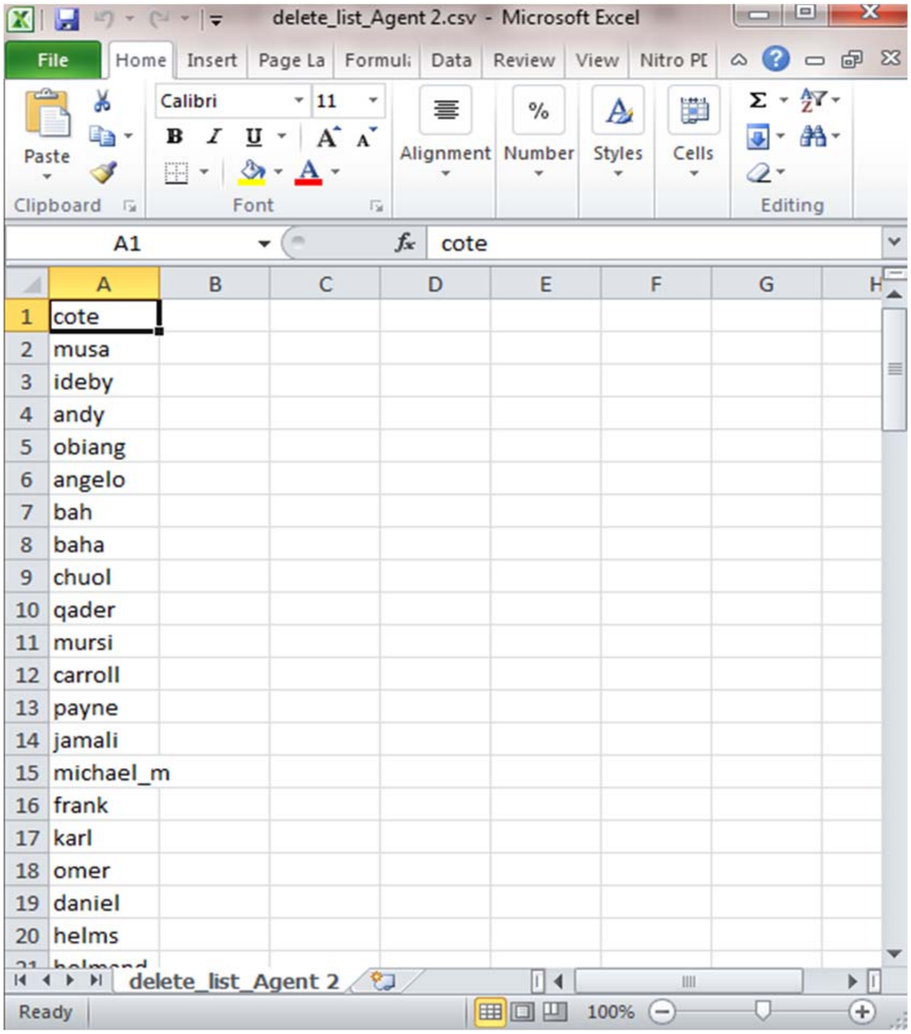
Delete lists allow the users to remove non-content bearing conjunctions, articles and other noise from texts. Delete lists can be created internally in AutoMap or externally in a text editor. The list itself is a text file that contains a list (one concept per line) of the words to be deleted from the text.

Note: Whether the users apply the Delete list(s) before or after applying a Thesaurus will depend on their individual needs.

7.1 Domain delete list

A domain delete list is one which contains concepts which need to be deleted and are project specific.

Note: #delete in the change format moves the CURRENT_CONCEPT to the domain delete list.



A1: cote

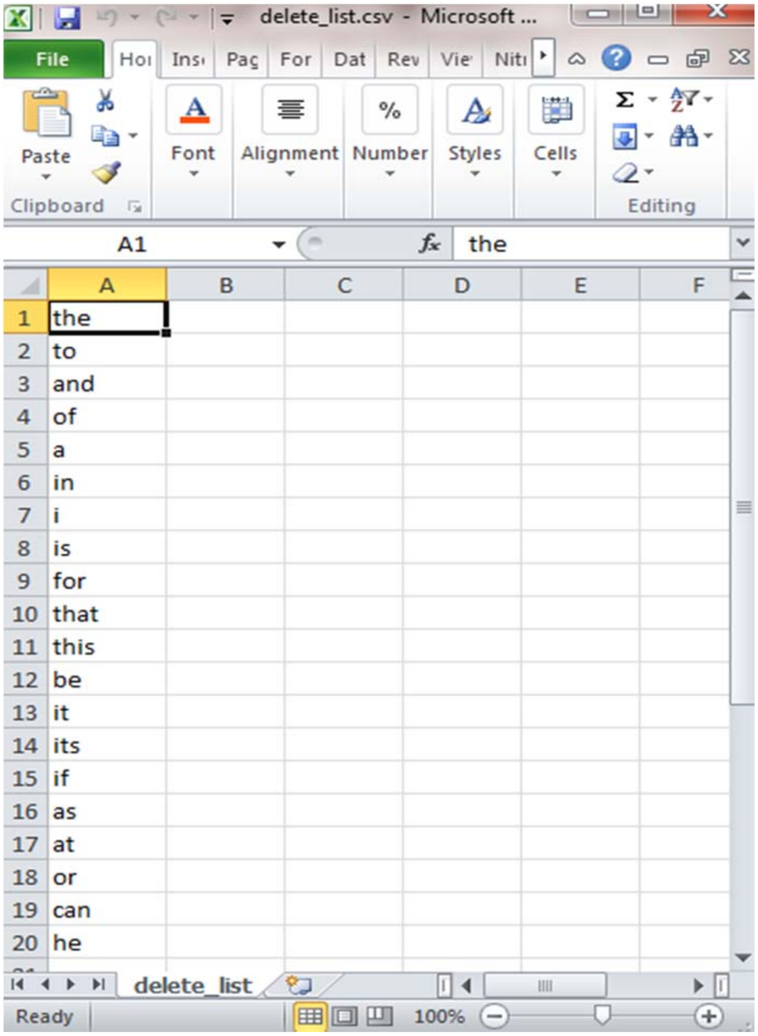
	A	B	C	D	E	F	G	H
1	cote							
2	musa							
3	ideby							
4	andy							
5	obiang							
6	angelo							
7	bah							
8	baha							
9	chuol							
10	qader							
11	mursi							
12	carroll							
13	payne							
14	jamali							
15	michael_m							
16	frank							
17	karl							
18	omer							
19	daniel							
20	helms							
21	helmand							

delete_list_Agent 2

Fig 10: Domain delete list

7.2 Universal delete list

A universal delete list contains concepts that need to be deleted and are generally comprised of the universal concepts like pronouns, months, numbers and other noise words.



A1 | *fx* the

	A	B	C	D	E	F
1	the					
2	to					
3	and					
4	of					
5	a					
6	in					
7	i					
8	is					
9	for					
10	that					
11	this					
12	be					
13	it					
14	its					
15	if					
16	as					
17	at					
18	or					
19	can					
20	he					

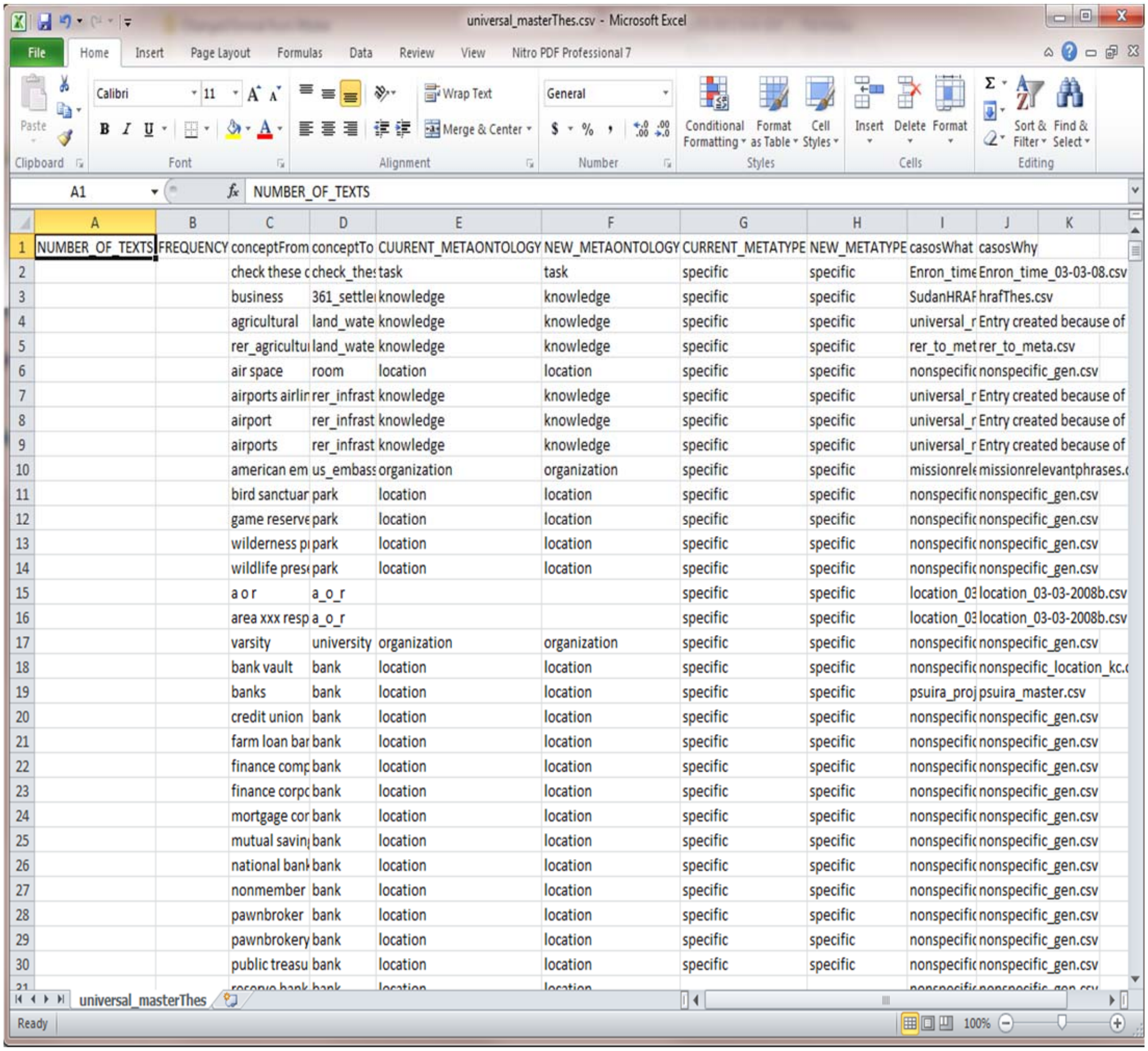
Fig 11: Universal delete list

8 Difference between the Domain Thesauri and the Universal Thesaurus

The product of the thesaurus creation process (i.e. the outcome from splitting and merging) can be classified into two major types of thesaurus files. The Universal/Master thesauri is intended for general application to all input text sets and the domain thesauri which contains concepts which are project specific and hence is used on specific text sets and not on others.

8.1 Universal Thesaurus

This is the major thesaurus which is created by the merge and split process. This thesaurus file contains all the concepts which are not specific to the projects (which are contained in domain thesaurus), the universal thesaurus contains all the common words, acronyms, common nouns, verbs, proverbs etc. Apart from these, the universal thesaurus also contains all the geographical locations and the organizations.



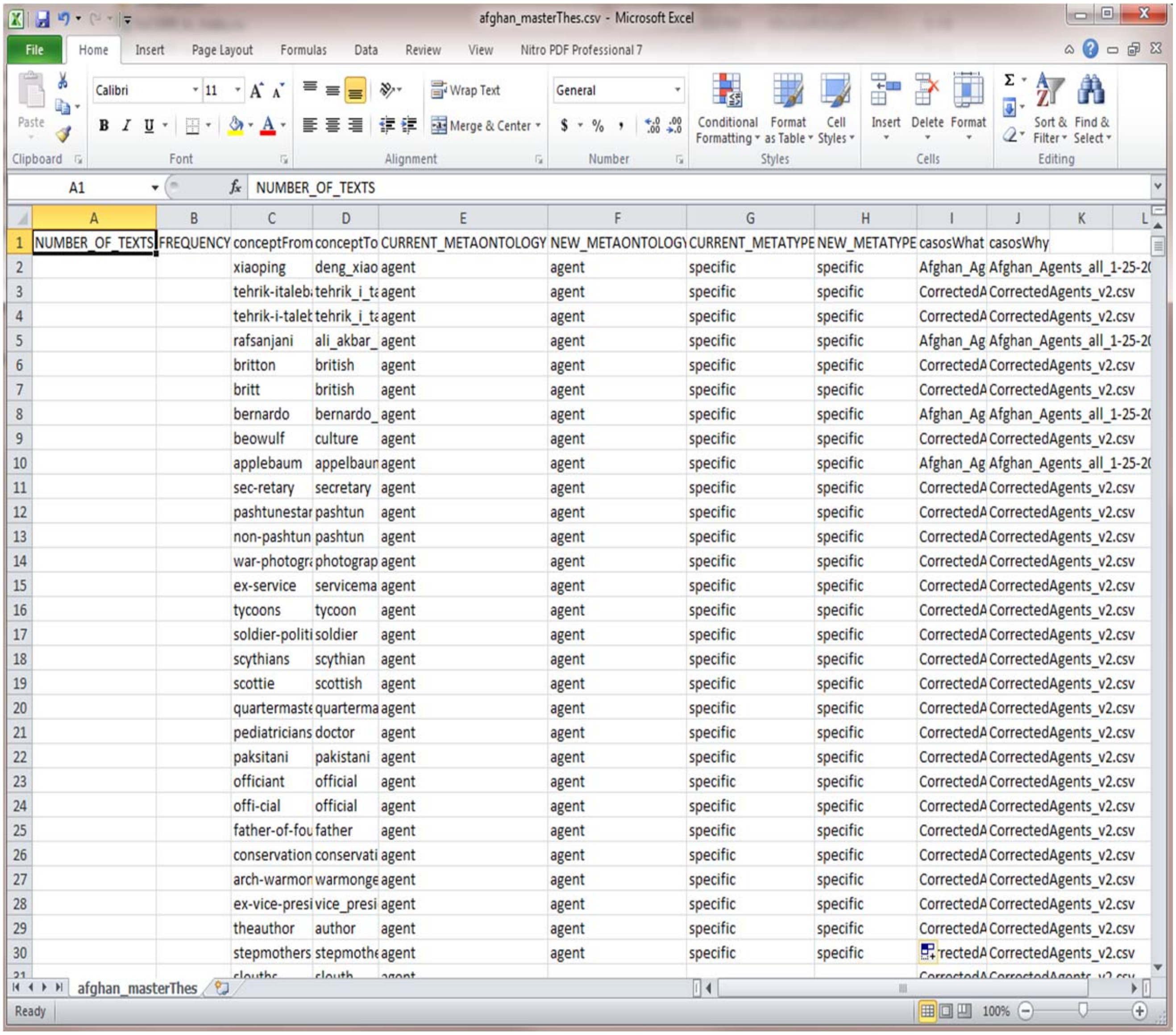
A1 NUMBER_OF_TEXTS

	A	B	C	D	E	F	G	H	I	J	K
1	NUMBER_OF_TEXTS	FREQUENCY	conceptFrom	conceptTo	CUURENT_METAONTOLOGY	NEW_METAONTOLOGY	CURRENT_METATYPE	NEW_METATYPE	casosWhat	casosWhy	
2			check these c	check_thes	task	task	specific	specific	Enron_time	Enron_time_03-03-08.csv	
3			business	361_settle	knowledge	knowledge	specific	specific	SudanHRAF	hrafThes.csv	
4			agricultural	land_wate	knowledge	knowledge	specific	specific	universal_r	Entry created because of	
5			rer_agricultu	land_wate	knowledge	knowledge	specific	specific	rer_to_met	rer_to_meta.csv	
6			air space	room	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
7			airports airlir	rer_infrast	knowledge	knowledge	specific	specific	universal_r	Entry created because of	
8			airport	rer_infrast	knowledge	knowledge	specific	specific	universal_r	Entry created because of	
9			airports	rer_infrast	knowledge	knowledge	specific	specific	universal_r	Entry created because of	
10			american em	us_embass	organization	organization	specific	specific	missionrele	missionrelevantphrases.c	
11			bird sanctuar	park	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
12			game reserve	park	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
13			wilderness p	park	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
14			wildlife prese	park	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
15			a o r	a_o_r			specific	specific	location_03	location_03-03-2008b.csv	
16			area xxx resp	a_o_r			specific	specific	location_03	location_03-03-2008b.csv	
17			varsity	university	organization	organization	specific	specific	nonspecific	nonspecific_gen.csv	
18			bank vault	bank	location	location	specific	specific	nonspecific	nonspecific_location_kc.c	
19			banks	bank	location	location	specific	specific	psuira_proj	psuira_master.csv	
20			credit union	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
21			farm loan bar	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
22			finance comp	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
23			finance corpc	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
24			mortgage cor	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
25			mutual savin	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
26			national bank	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
27			nonmember	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
28			pawnbroker	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
29			pawnbrokery	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
30			public treasu	bank	location	location	specific	specific	nonspecific	nonspecific_gen.csv	
31			reserve bank	bank	location	location			nonspecific	nonspecific_gen.csv	

Fig 12: Universal master thesaurus (universal_masterThes.csv)

8.2 Domain Thesauri

The domain thesauri are project specific thesauri produced by the merge and split process. According to our approach, the domain thesaurus contains only agents which are specific to the project. The approach used to differentiate between the project specific agents and the universal agents is that the project specific agents were considered to be the ones which were one word names, i.e. no white spaces were present.



fx NUMBER_OF_TEXTS

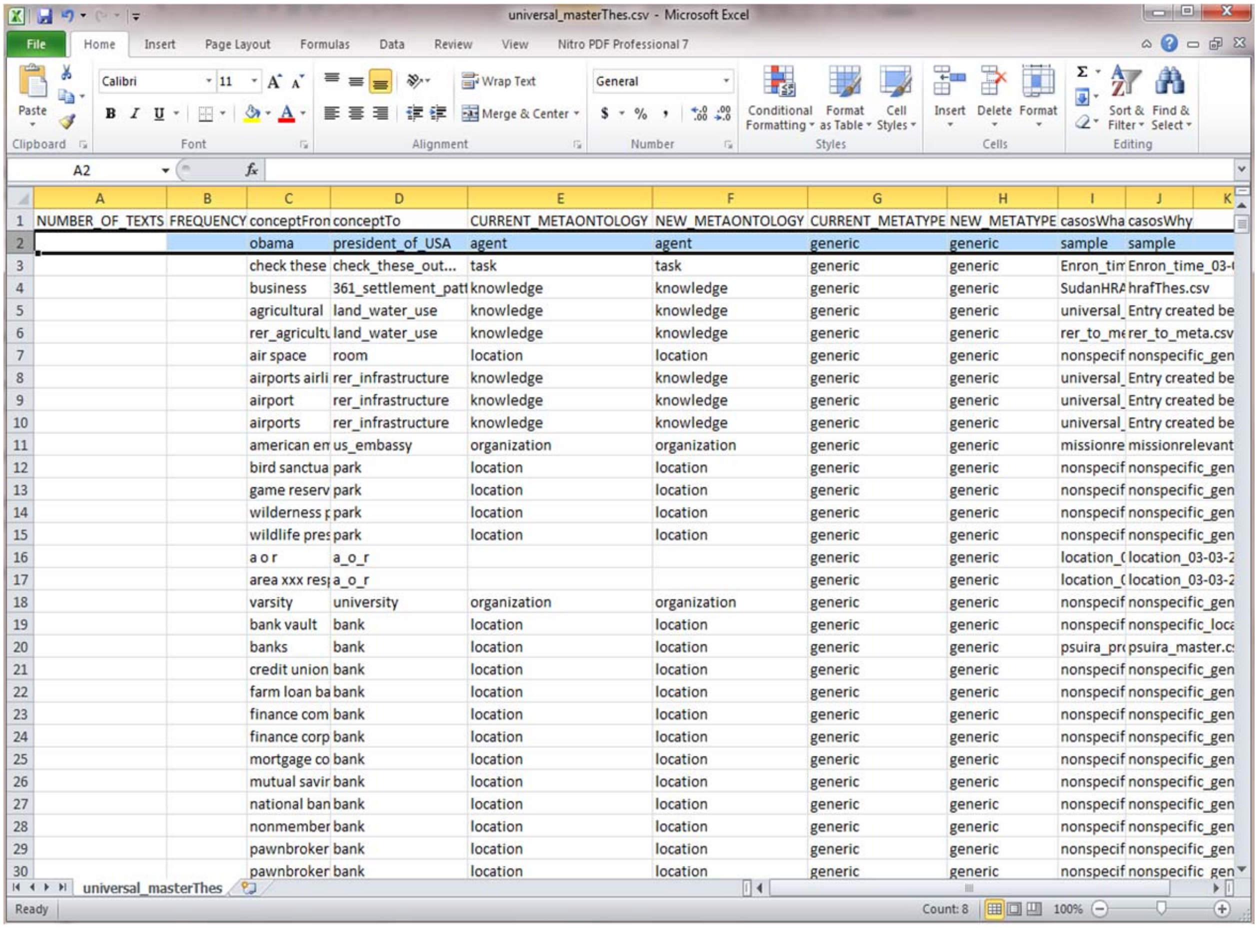
	A	B	C	D	E	F	G	H	I	J	K	L
1	NUMBER_OF_TEXTS	FREQUENCY	conceptFrom	conceptTo	CURRENT_METAONTOLOGY	NEW_METAONTOLOGY	CURRENT_METATYPE	NEW_METATYPE	casosWhat	casosWhy		
2			xiaoping	deng_xiao	agent	agent	specific	specific	Afghan_Ag	Afghan_Agents_all_1-25-20		
3			tehrik-italeb	tehrik_i_ta	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
4			tehrik-i-talet	tehrik_i_ta	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
5			rafsanjani	ali_akbar_	agent	agent	specific	specific	Afghan_Ag	Afghan_Agents_all_1-25-20		
6			britton	british	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
7			britt	british	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
8			bernardo	bernardo_	agent	agent	specific	specific	Afghan_Ag	Afghan_Agents_all_1-25-20		
9			beowulf	culture	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
10			applebaum	appelbaun	agent	agent	specific	specific	Afghan_Ag	Afghan_Agents_all_1-25-20		
11			sec-retary	secretary	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
12			pashtunestar	pashtun	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
13			non-pashtun	pashtun	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
14			war-photogra	photograp	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
15			ex-service	servicema	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
16			tycoons	tycoon	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
17			soldier-politi	soldier	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
18			scythians	scythian	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
19			scottie	scottish	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
20			quartermaste	quarterma	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
21			pediatricians	doctor	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
22			paksitani	pakistani	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
23			officiant	official	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
24			offi-cial	official	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
25			father-of-fou	father	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
26			conservation	conservati	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
27			arch-warmon	warmonge	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
28			ex-vice-presi	vice_presi	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
29			theauthor	author	agent	agent	specific	specific	CorrectedA	CorrectedAgents_v2.csv		
30			stepmothers	stepmothe	agent	agent	specific	specific	rectedA	CorrectedAgents_v2.csv		
31			clouths	clouth	agent				CorrectedA	CorrectedAgents_v2.csv		

Fig 13: Domain thesaurus (afghan_masterThes.csv)

9 Dominance of Domain Thesauri

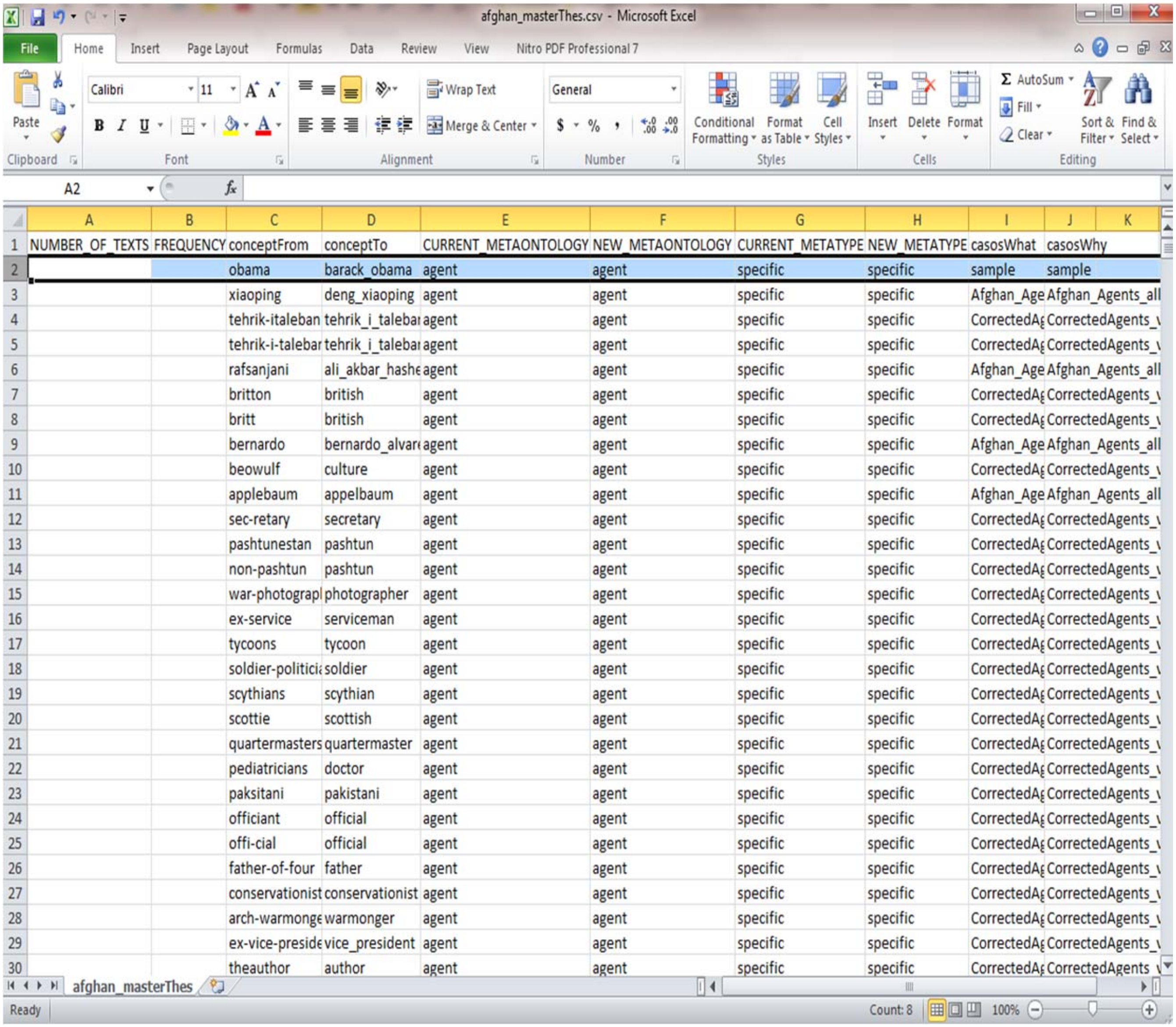
Apart from the difference in the content and the data contained by the two thesaurus files, these two types of thesauri differ in the precedence as well. The mapping in the domain thesauri takes precedence over the mapping in the universal thesaurus when both types of thesaurus files are used when performing data analysis by generating a Meta-network out of a corpus of raw text data.

Example: Consider a data set which needs to be analyzed by generating the Dynetml (network) files using AutoMap. In order to generate this network we provide the two thesaurus files (the domain thesaurus and the universal thesaurus). Now, assume that the word **Obama** appears in the body of text for analysis. AutoMap references the thesauri files to map the concept, and the concept **Obama** is mapped to barack_obama in the Domain thesaurus and in the universal thesauri, it is mapped as president_of_USA. AutoMap applies mapping which is in the domain thesaurus giving preference to the domain thesaurus over the universal thesaurus. Thus, we say that the Domain thesaurus has precedence over the Universal thesaurus.



	A	B	C	D	E	F	G	H	I	J
1	NUMBER_OF_TEXTS	FREQUENCY	conceptFron	conceptTo	CURRENT_METAONTOLOGY	NEW_METAONTOLOGY	CURRENT_METATYPE	NEW_METATYPE	casosWha	casosWhy
2			obama	president_of_USA	agent	agent	generic	generic	sample	sample
3			check these	check_these_out...	task	task	generic	generic	Enron_tim	Enron_time_03-(
4			business	361_settlement_patt	knowledge	knowledge	generic	generic	SudanHRA	hrafThes.csv
5			agricultural	land_water_use	knowledge	knowledge	generic	generic	universal_	Entry created be
6			rer_agricultu	land_water_use	knowledge	knowledge	generic	generic	rer_to_me	rer_to_meta.csv
7			air space	room	location	location	generic	generic	nonspecif	nonspecific_gen
8			airports airli	rer_infrastructure	knowledge	knowledge	generic	generic	universal_	Entry created be
9			airport	rer_infrastructure	knowledge	knowledge	generic	generic	universal_	Entry created be
10			airports	rer_infrastructure	knowledge	knowledge	generic	generic	universal_	Entry created be
11			american en	us_embassy	organization	organization	generic	generic	missionre	missionrelevant
12			bird sanctua	park	location	location	generic	generic	nonspecif	nonspecific_gen
13			game reserv	park	location	location	generic	generic	nonspecif	nonspecific_gen
14			wilderness p	park	location	location	generic	generic	nonspecif	nonspecific_gen
15			wildlife pres	park	location	location	generic	generic	nonspecif	nonspecific_gen
16			a o r	a_o_r			generic	generic	location_(	location_03-03-2
17			area xxx resp	a_o_r			generic	generic	location_(	location_03-03-2
18			varsity	university	organization	organization	generic	generic	nonspecif	nonspecific_gen
19			bank vault	bank	location	location	generic	generic	nonspecif	nonspecific_loca
20			banks	bank	location	location	generic	generic	psuira_pr(	psuira_master.c:
21			credit union	bank	location	location	generic	generic	nonspecif	nonspecific_gen
22			farm loan ba	bank	location	location	generic	generic	nonspecif	nonspecific_gen
23			finance com	bank	location	location	generic	generic	nonspecif	nonspecific_gen
24			finance corp	bank	location	location	generic	generic	nonspecif	nonspecific_gen
25			mortgage co	bank	location	location	generic	generic	nonspecif	nonspecific_gen
26			mutual savir	bank	location	location	generic	generic	nonspecif	nonspecific_gen
27			national ban	bank	location	location	generic	generic	nonspecif	nonspecific_gen
28			nonmember	bank	location	location	generic	generic	nonspecif	nonspecific_gen
29			pawnbroker	bank	location	location	generic	generic	nonspecif	nonspecific_gen
30			pawnbroker	bank	location	location	generic	generic	nonspecif	nonspecific_gen

Fig 14: Universal master thesaurus (universal_masterThes.csv)



	A	B	C	D	E	F	G	H	I	J
1	NUMBER_OF_TEXTS	FREQUENCY	conceptFrom	conceptTo	CURRENT_METAONTOLOGY	NEW_METAONTOLOGY	CURRENT_METATYPE	NEW_METATYPE	casosWhat	casosWhy
2			obama	barack_obama	agent	agent	specific	specific	sample	sample
3			xiaoping	deng_xiaoping	agent	agent	specific	specific	Afghan_Age	Afghan_Agents_all
4			tehrik-italeban	tehrik_i_talebar	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
5			tehrik-i-talebar	tehrik_i_talebar	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
6			rafsanjani	ali_akbar_hashe	agent	agent	specific	specific	Afghan_Age	Afghan_Agents_all
7			britton	british	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
8			britt	british	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
9			bernardo	bernardo_alvare	agent	agent	specific	specific	Afghan_Age	Afghan_Agents_all
10			beowulf	culture	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
11			applebaum	appelbaum	agent	agent	specific	specific	Afghan_Age	Afghan_Agents_all
12			sec-retary	secretary	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
13			pashtunestan	pashtun	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
14			non-pashtun	pashtun	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
15			war-photograpl	photographer	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
16			ex-service	serviceman	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
17			tycoons	tycoon	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
18			soldier-politici	soldier	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
19			scythians	scythian	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
20			scottie	scottish	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
21			quartermasters	quartermaster	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
22			pediatricians	doctor	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
23			paksitani	pakistani	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
24			officiant	official	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
25			offi-cial	official	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
26			father-of-four	father	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
27			conservationist	conservationist	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
28			arch-warmonge	warmonger	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
29			ex-vice-preside	vice_president	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v
30			theauthor	author	agent	agent	specific	specific	CorrectedAg	CorrectedAgents_v

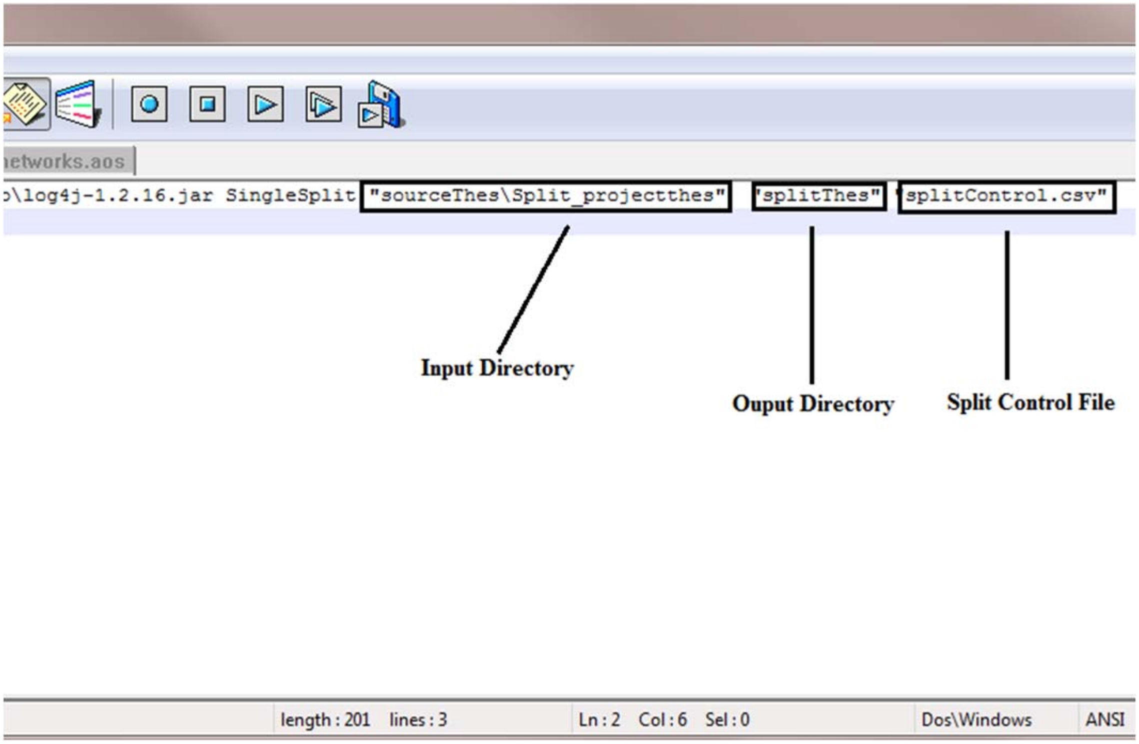
Fig 15: Domain thesaurus (afghan_masterThes.csv)

10 The Split Process

The development copy of AutoMap contains the required files and utilities for thesaurus split and merge. Typically a user will check out a copy of AutoMap from the Subversion repository prior to applying the process. At present this process is not performed using the general distribution copy of AutoMap

The split process used during thesaurus preparation uses these steps:

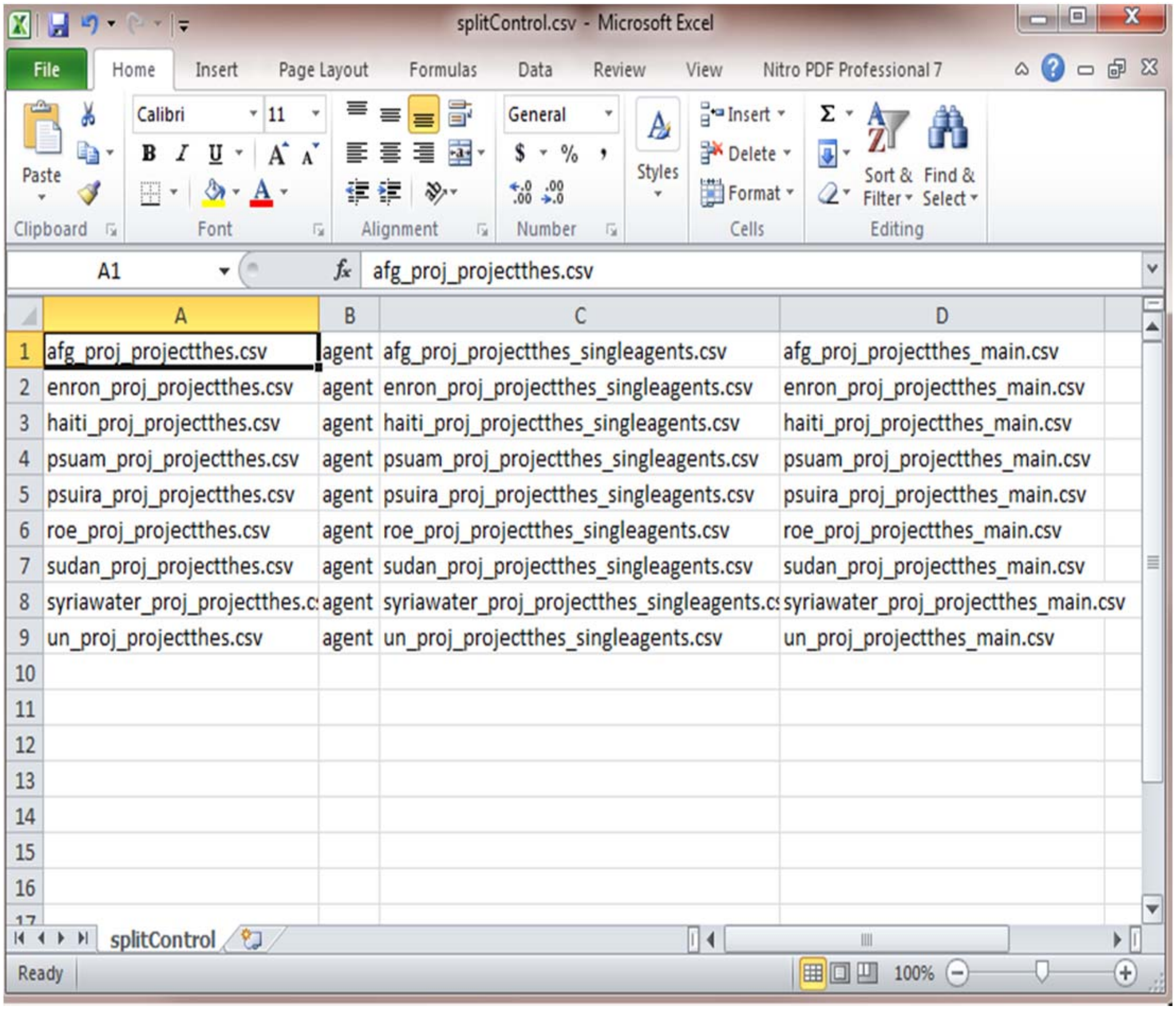
1. Open the **splitThes.bat** in notepad in order to explore the file list for the Split process.
2. Run the split process using **SplitThes.bat** which is located in Automap3\usr\etc directory.



```
>\log4j-1.2.16.jar SingleSplit "sourceThes\Split_projectthes" 'splitThes" 'splitControl.csv"
```

Fig 16: Screen shot of the file **splitThes.bat**

3. The input directory should contain all the files which need to be split and should be listed in the file **splitcontrol.csv**
4. The output directory **splitThes** will contain all the files after the split process is completed.
5. The split control file generally has to be modified by the users to process the specific thesauri needed by the user. The column A contains the list of all the files which need to be split. The column B contains the Meta Ontology value on which the file needs to be split ("agent" has been used in this example). Column C contains the name of the file where the split entries should be written and the column D contains the name of the file where all other entries need to be written.



A1 | afg_proj_projectthes.csv

	A	B	C	D
1	afg_proj_projectthes.csv	agent	afg_proj_projectthes_singleagents.csv	afg_proj_projectthes_main.csv
2	enron_proj_projectthes.csv	agent	enron_proj_projectthes_singleagents.csv	enron_proj_projectthes_main.csv
3	haiti_proj_projectthes.csv	agent	haiti_proj_projectthes_singleagents.csv	haiti_proj_projectthes_main.csv
4	psuam_proj_projectthes.csv	agent	psuam_proj_projectthes_singleagents.csv	psuam_proj_projectthes_main.csv
5	psuira_proj_projectthes.csv	agent	psuira_proj_projectthes_singleagents.csv	psuira_proj_projectthes_main.csv
6	roe_proj_projectthes.csv	agent	roe_proj_projectthes_singleagents.csv	roe_proj_projectthes_main.csv
7	sudan_proj_projectthes.csv	agent	sudan_proj_projectthes_singleagents.csv	sudan_proj_projectthes_main.csv
8	syriawater_proj_projectthes.c	agent	syriawater_proj_projectthes_singleagents.c	syriawater_proj_projectthes_main.csv
9	un_proj_projectthes.csv	agent	un_proj_projectthes_singleagents.csv	un_proj_projectthes_main.csv
10				
11				
12				
13				
14				
15				
16				

Fig 17: Screen shot for the file **splitcontrol.csv**

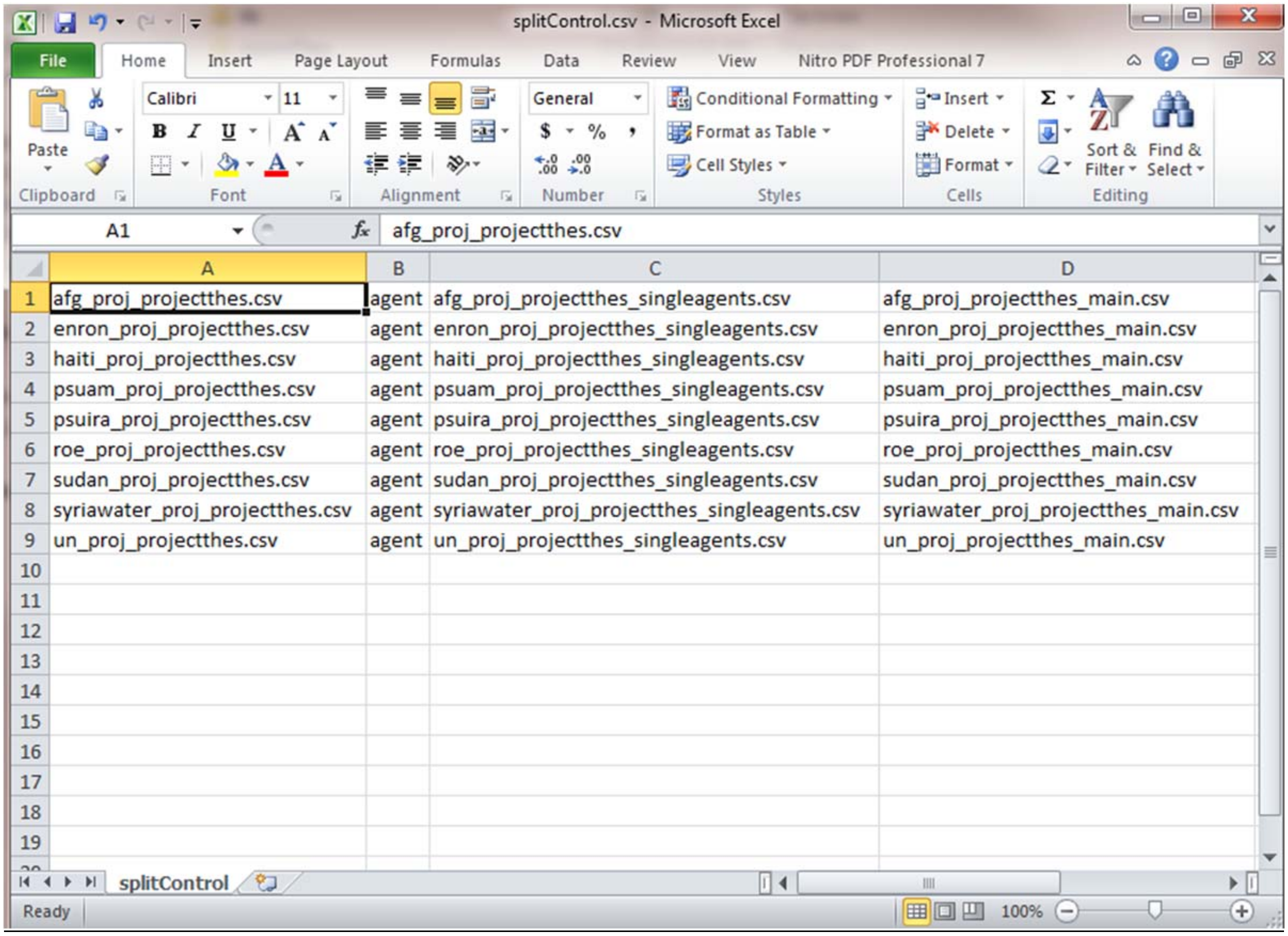
Note: The format used shown above should be strictly used as the process may break down if there are any modifications. The files which need to be merged should be in the master thesauri format.

11 Split Routine

The split routine helps in the splitting of the entries in a file on the basis of an algorithm. The split process used in this report splits the files if the meta ontology of the entry in a file is an AGENT and the word is a single word (without any white spaces) as discussed in the previous section.

Steps involved in the code of the split routine:

1. Opens the file **splitcontrol.csv** and processes the input line by line.

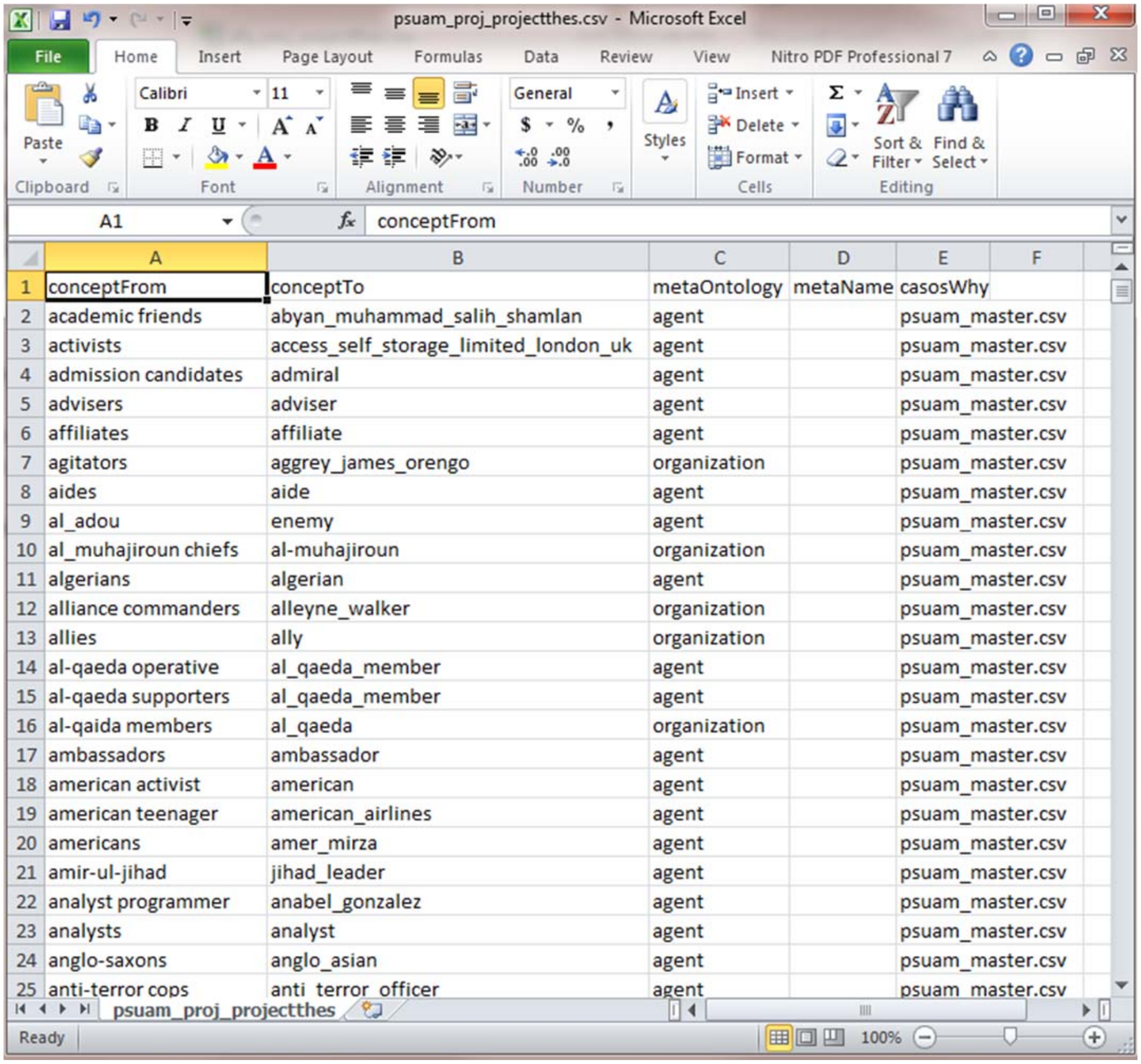


A1: afg_proj_projectthes.csv

	A	B	C	D
1	afg_proj_projectthes.csv	agent	afg_proj_projectthes_singleagents.csv	afg_proj_projectthes_main.csv
2	enron_proj_projectthes.csv	agent	enron_proj_projectthes_singleagents.csv	enron_proj_projectthes_main.csv
3	haiti_proj_projectthes.csv	agent	haiti_proj_projectthes_singleagents.csv	haiti_proj_projectthes_main.csv
4	psuam_proj_projectthes.csv	agent	psuam_proj_projectthes_singleagents.csv	psuam_proj_projectthes_main.csv
5	psuira_proj_projectthes.csv	agent	psuira_proj_projectthes_singleagents.csv	psuira_proj_projectthes_main.csv
6	roe_proj_projectthes.csv	agent	roe_proj_projectthes_singleagents.csv	roe_proj_projectthes_main.csv
7	sudan_proj_projectthes.csv	agent	sudan_proj_projectthes_singleagents.csv	sudan_proj_projectthes_main.csv
8	syriawater_proj_projectthes.csv	agent	syriawater_proj_projectthes_singleagents.csv	syriawater_proj_projectthes_main.csv
9	un_proj_projectthes.csv	agent	un_proj_projectthes_singleagents.csv	un_proj_projectthes_main.csv

Fig 18: Format for the file **splitcontrol.csv**

2. Reads the entry in column A and opens the named file.



	A	B	C	D	E	F
1	conceptFrom	conceptTo	metaOntology	metaName	casosWhy	
2	academic friends	abyan_muhammad_salih_shamlan	agent		psuam_master.csv	
3	activists	access_self_storage_limited_london_uk	agent		psuam_master.csv	
4	admission candidates	admiral	agent		psuam_master.csv	
5	advisers	adviser	agent		psuam_master.csv	
6	affiliates	affiliate	agent		psuam_master.csv	
7	agitators	aggrey_james_orengo	organization		psuam_master.csv	
8	aides	aide	agent		psuam_master.csv	
9	al_adou	enemy	agent		psuam_master.csv	
10	al_muhajiroun chiefs	al-muhajiroun	organization		psuam_master.csv	
11	algerians	algerian	agent		psuam_master.csv	
12	alliance commanders	alleyne_walker	organization		psuam_master.csv	
13	allies	ally	organization		psuam_master.csv	
14	al-qaeda operative	al_qaeda_member	agent		psuam_master.csv	
15	al-qaeda supporters	al_qaeda_member	agent		psuam_master.csv	
16	al-qaida members	al_qaeda	organization		psuam_master.csv	
17	ambassadors	ambassador	agent		psuam_master.csv	
18	american activist	american	agent		psuam_master.csv	
19	american teenager	american_airlines	agent		psuam_master.csv	
20	americans	amer_mirza	agent		psuam_master.csv	
21	amir-ul-jihad	jihad_leader	agent		psuam_master.csv	
22	analyst programmer	anabel_gonzalez	agent		psuam_master.csv	
23	analysts	analyst	agent		psuam_master.csv	
24	anglo-saxons	anglo_asian	agent		psuam_master.csv	
25	anti-terror cops	anti_terror_officer	agent		psuam_master.csv	

Fig 19: Format for the file in the column A

3. Once the file is opened, the contents are scanned line by line (note however that the header line in the input thesaurus is skipped).
4. The first entry "academic friends" has a white space, so this entry is written to the file name in column D of the **psuam_proj_projectthes.csv** which is the file **psuam_proj_projectthes_main.csv**.
5. The second entry "activists" does not have any white space and the corresponding Meta ontology is also agents, so this entry goes to the file named in column C of the **splitcontol.csv**, i.e. **psuam_proj_projectthes_singleagents.csv**.

Note: The split routine can be modified according to the needs of the user by changing the column B of **splitcontrol.csv** to some other Meta ontology.

12 Algorithm for the Split Routine

This section describes the algorithm for the split process. The algorithm has been simplified as much as possible for the convenience of the reader.

Algorithm – Split routine.

```
Input (splitcontrol.csv)
Row1=0;

While (row1 does not reach the end of file)
{
    Row2=0;
    Read (column A of splitcontrol.csv)
    Open (thesaurus-file corresponding to column A of splitcontrol.csv)

        While (the Row2 does not reach the end of thesaurus-file)
        {
            Read the conceptFrom column;
            If (the entry contains a white space character OR underscores character)
           Write row contents to the thesaurus_file_main.csv

           Else
           Write row contents to the thesaurus_file_singleagents.csv

          Row2++;
       }

Row1++;
}

Close (thesaurus_file.csv)
```

13 The Merge Process

The merge routine is not a simple additive process of summing all the concepts from the input files into one thesaurus; it involves some computing operations as well, after we have created a merged thesaurus. These operations are Remove conflicts, Detect cycles, and Depth first search. The merge process used uses the following series of steps:

1. **Remove conflicts:** After all the input files are merged into one thesaurus file, the function checks for duplicate entries in the conceptFrom column. If an entry is found more than once, it deletes all other occurrences except the first one.

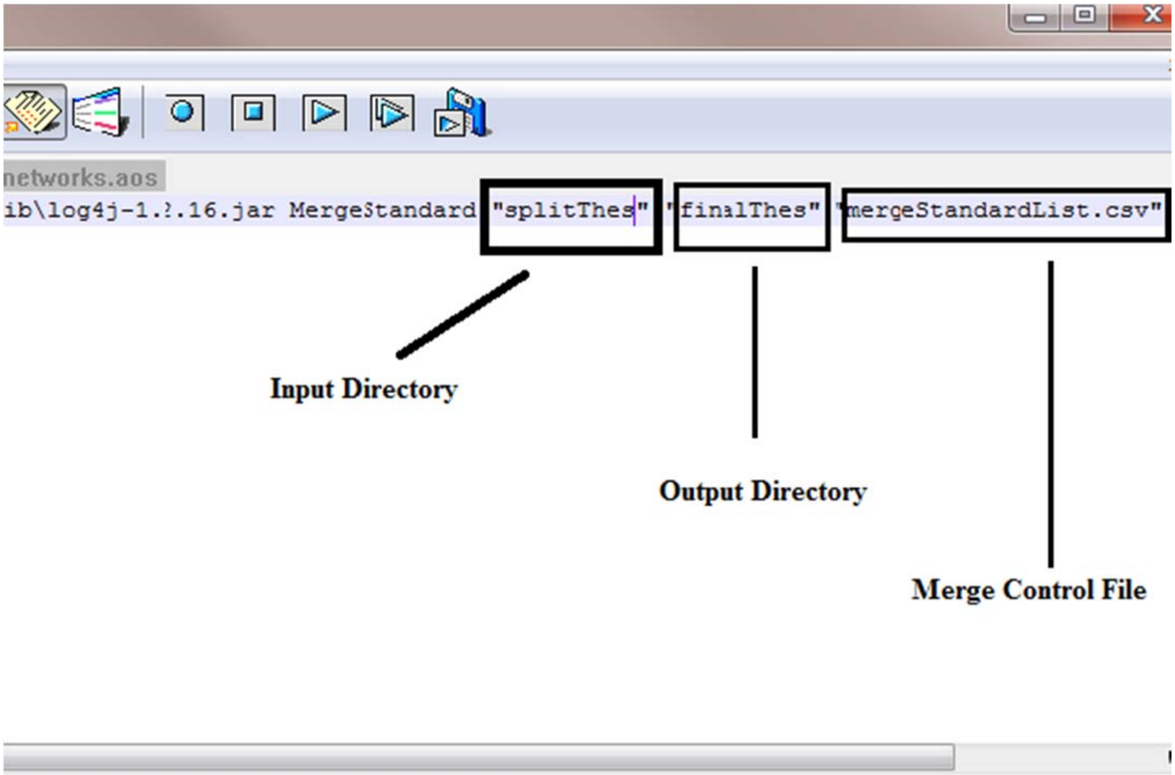
2. **Detect Cycles:** This function checks if any cycle is created due to merging all the input files. If a cycle of concepts is created, it just maps the concept to itself. For example: Consider that India is mapped to country, country to world, world to tradition, tradition to India. This means that consequently India is mapped to itself, so this function simply matches India to India.

3. **Depth first search**: This function essentially checks all the chain mapping involved in the merged file. Once the end point of chain mapping is found, it simply creates a mapping from the head to the tail. For example: AutoMap is mapped to Network Analysis, Network Analysis mapped to CASOS, CASOS mapped to CMU. So, this function simply maps AutoMap to CMU.

Note: The order in which these functions appear in the merge algorithm is also significant.

The steps involved in the Merge process are:

1. Open the **mergeStandard.bat** in notepad in order to see the list of files and directories employed in the Merge process. **(MergeStandard.bat is locat**ed in AutoMap3\usr\etc directory.)

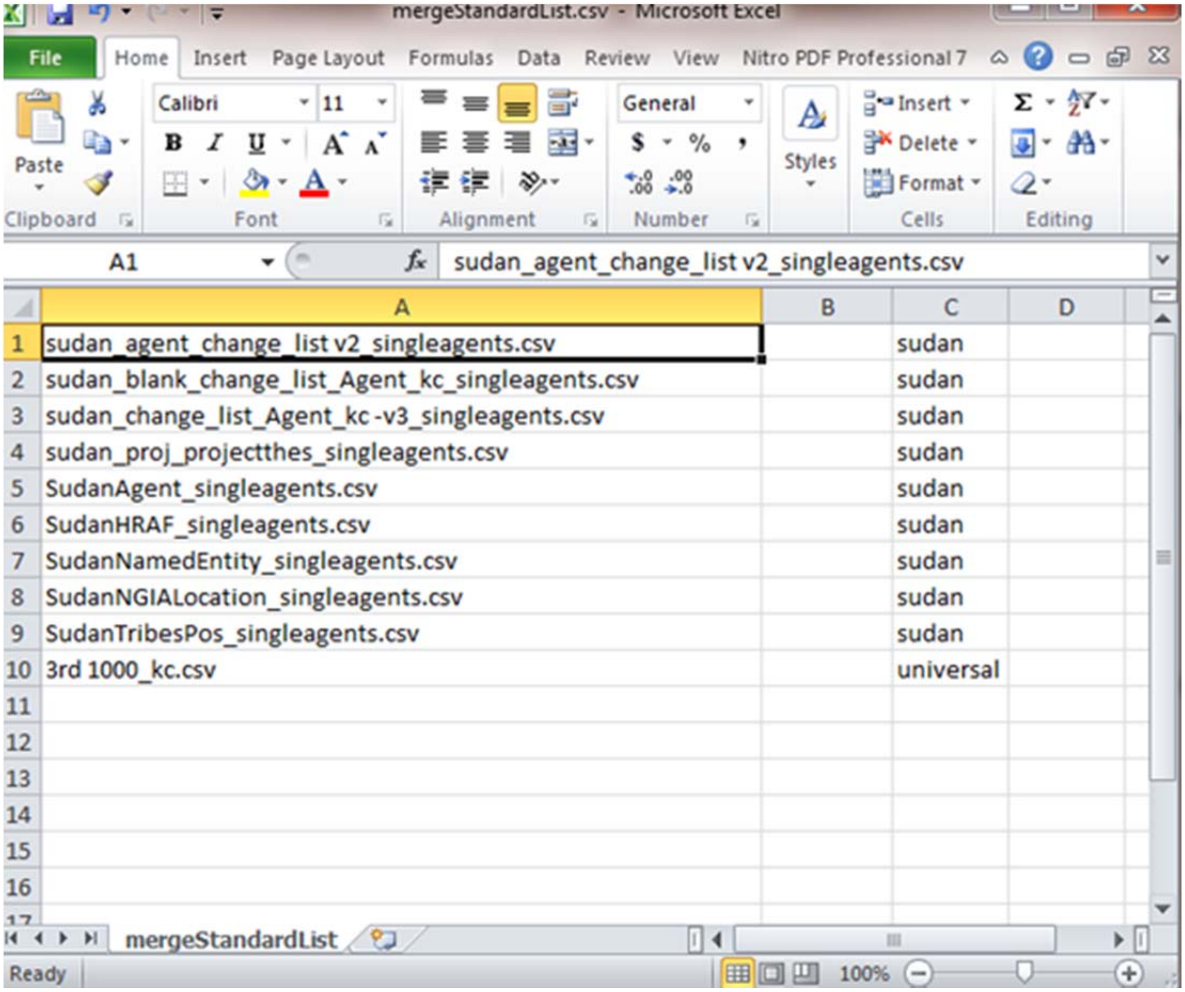


```
ib\log4j-1.2.16.jar MergeStandard "splitThes" "finalThes" "mergeStandardList.csv"
```

Fig 20: Screen shot for the file **mergeStandard.bat**

2. **The input dire**ctory "**splitThes**" should contain all the files which need to be split. No sub directories are allowed in this directory.

3. The merge control file, **mergeStandard.csv** controls the merge process. In this file, all the files which need **to** be merged are listed as well as the category in which they need to be merged is listed. For example: Universal, Sudan, Afghan etc.



A1 *fx* sudan_agent_change_list v2_singleagents.csv

	A	B	C	D
1	sudan_agent_change_list v2_singleagents.csv		sudan	
2	sudan_blank_change_list_Agent_kc_singleagents.csv		sudan	
3	sudan_change_list_Agent_kc -v3_singleagents.csv		sudan	
4	sudan_proj_projectthes_singleagents.csv		sudan	
5	SudanAgent_singleagents.csv		sudan	
6	SudanHRAF_singleagents.csv		sudan	
7	SudanNamedEntity_singleagents.csv		sudan	
8	SudanNGIALocation_singleagents.csv		sudan	
9	SudanTribesPos_singleagents.csv		sudan	
10	3rd 1000_kc.csv		universal	
11				
12				
13				
14				
15				
16				

Fig 21: Screen shot for the file **mergeStandardList.csv**

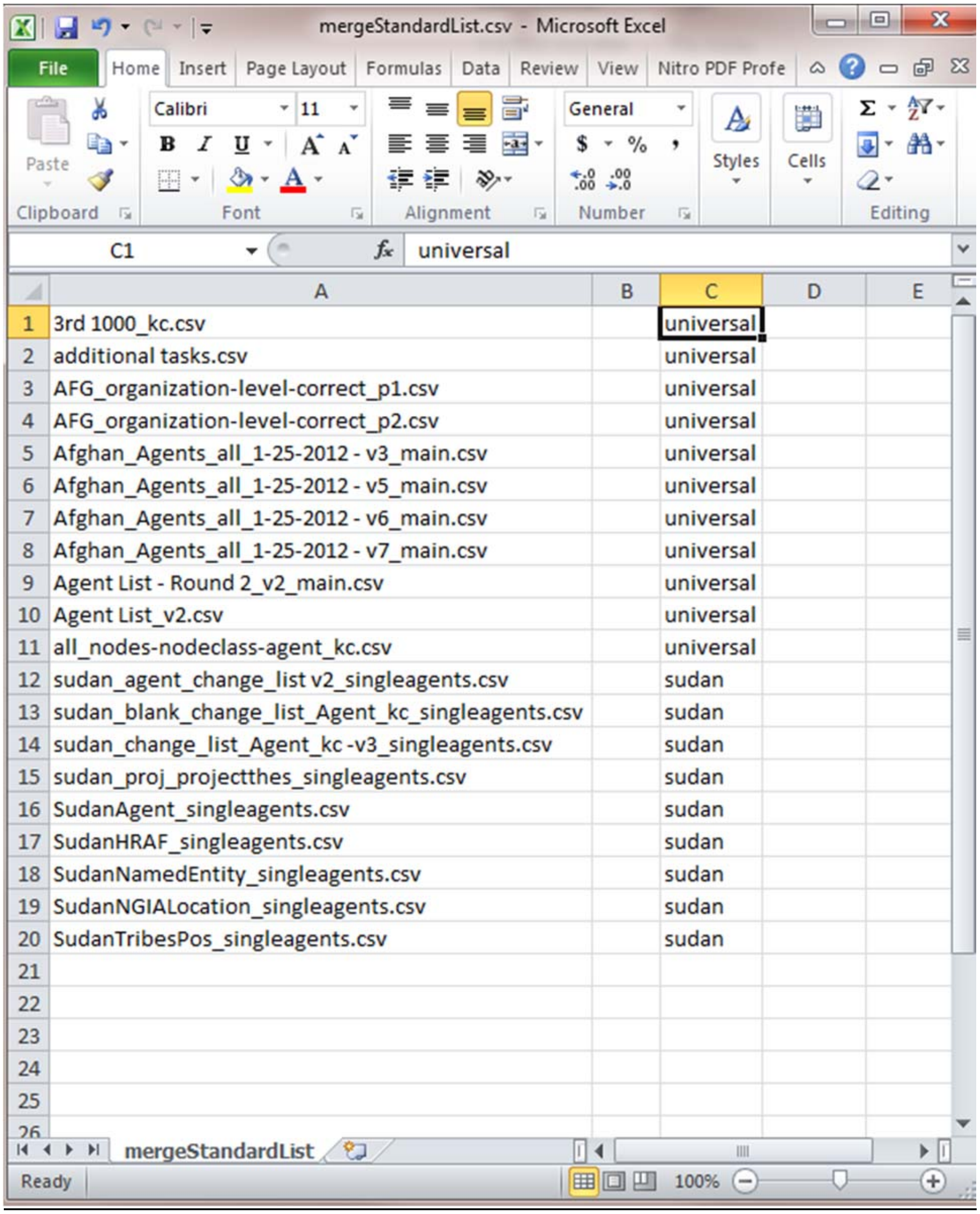
Note: The Column B of the mergeStandardList.csv can be left as it is not used by merge routine.

14 Merge Routine

The merge routine is used to the merge a group of thesaurus files in a single merged thesaurus file on the basis of an algorithm.

Steps involved in the code of the merge routine are as follows:

1. Opens the file **mergeStandardList.csv** and scans the input line by line.

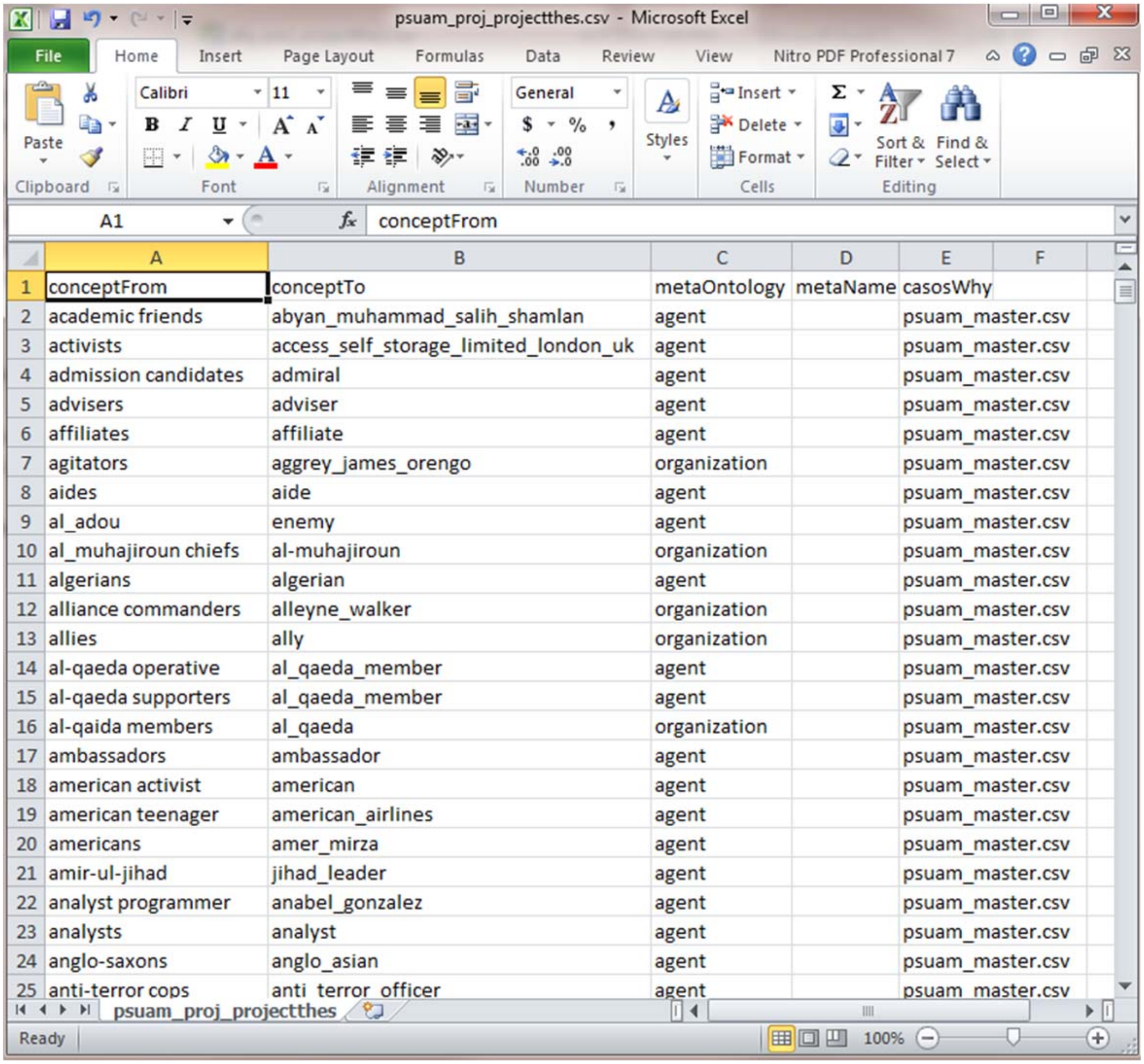


C1 = universal

	A	B	C	D	E
1	3rd 1000_kc.csv		universal		
2	additional tasks.csv		universal		
3	AFG_organization-level-correct_p1.csv		universal		
4	AFG_organization-level-correct_p2.csv		universal		
5	Afghan_Agents_all_1-25-2012 - v3_main.csv		universal		
6	Afghan_Agents_all_1-25-2012 - v5_main.csv		universal		
7	Afghan_Agents_all_1-25-2012 - v6_main.csv		universal		
8	Afghan_Agents_all_1-25-2012 - v7_main.csv		universal		
9	Agent List - Round 2_v2_main.csv		universal		
10	Agent List_v2.csv		universal		
11	all_nodes-nodeclass-agent_kc.csv		universal		
12	sudan_agent_change_list v2_singleagents.csv		sudan		
13	sudan_blank_change_list_Agent_kc_singleagents.csv		sudan		
14	sudan_change_list_Agent_kc -v3_singleagents.csv		sudan		
15	sudan_proj_projectthes_singleagents.csv		sudan		
16	SudanAgent_singleagents.csv		sudan		
17	SudanHRAF_singleagents.csv		sudan		
18	SudanNamedEntity_singleagents.csv		sudan		
19	SudanNGIALocation_singleagents.csv		sudan		
20	SudanTribesPos_singleagents.csv		sudan		
21					
22					
23					
24					
25					

Fig 22: Format for the file **mergestandardlist.csv**

2. Reads the entry **corresponding** to column A and opens the named thesaurus file.



	A	B	C	D	E	F
1	conceptFrom	conceptTo	metaOntology	metaName	casosWhy	
2	academic friends	abyan_muhammad_salih_shamlan	agent		psuam_master.csv	
3	activists	access_self_storage_limited_london_uk	agent		psuam_master.csv	
4	admission candidates	admiral	agent		psuam_master.csv	
5	advisers	adviser	agent		psuam_master.csv	
6	affiliates	affiliate	agent		psuam_master.csv	
7	agitators	aggrey_james_orengo	organization		psuam_master.csv	
8	aides	aide	agent		psuam_master.csv	
9	al_adou	enemy	agent		psuam_master.csv	
10	al_muhajiroun chiefs	al-muhajiroun	organization		psuam_master.csv	
11	algerians	algerian	agent		psuam_master.csv	
12	alliance commanders	alleyne_walker	organization		psuam_master.csv	
13	allies	ally	organization		psuam_master.csv	
14	al-qaeda operative	al_qaeda_member	agent		psuam_master.csv	
15	al-qaeda supporters	al_qaeda_member	agent		psuam_master.csv	
16	al-qaida members	al_qaeda	organization		psuam_master.csv	
17	ambassadors	ambassador	agent		psuam_master.csv	
18	american activist	american	agent		psuam_master.csv	
19	american teenager	american_airlines	agent		psuam_master.csv	
20	americans	amer_mirza	agent		psuam_master.csv	
21	amir-ul-jihad	jihad_leader	agent		psuam_master.csv	
22	analyst programmer	anabel_gonzalez	agent		psuam_master.csv	
23	analysts	analyst	agent		psuam_master.csv	
24	anglo-saxons	anglo_asian	agent		psuam_master.csv	
25	anti-terror cops	anti_terror_officer	agent		psuam_master.csv	

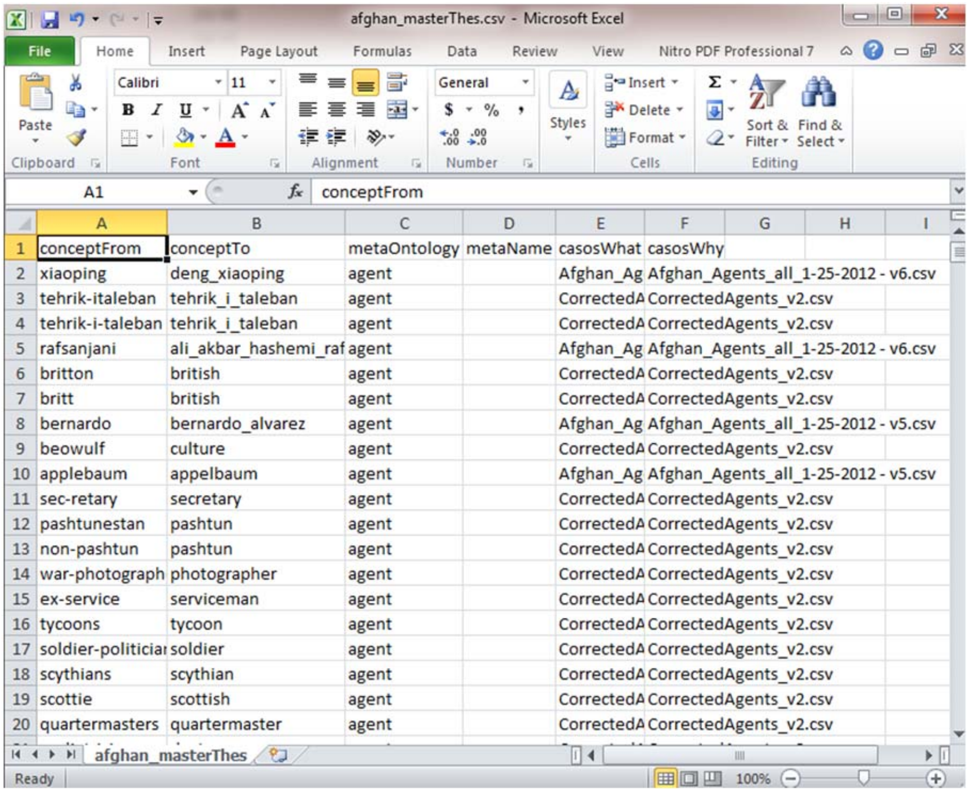
Fig 23: Format for the file in the column A

3. Column C of the entry line contains "universal", so all the entries in **3rd 1000_kc.csv** are merged into the file **universal_masterThes.csv** thesaurus file (the output name is <Column C contents> + "_masterThes.csv").

4. In the 12th line the file **sudan_agent_change_list_Agent_kc_singleagents.csv** has 'Sudan' in Column C, so all the entries are copied to the **sudan_masterThes.csv**

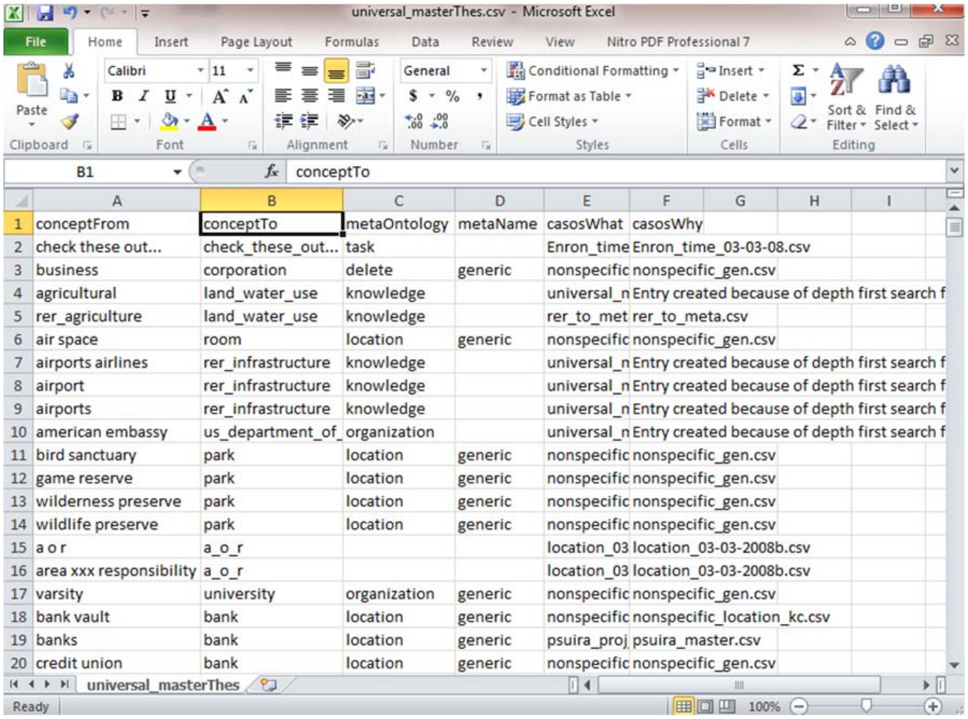
15 Algorithm for the Merge Routine

```
Open (mergestandardlist.csv)
Row1=0;

While (row1 does not reach the end of file)
{
      Row2=0;
      Grouping= Read (Column C of the file mergestandardlist.csv)
      Read (Column A of mergestandardlist.csv)
      Open (file corresponding to Column A of splitcontrol.csv)

       Open (grouping+"_masterThes.csv")
           While (the Row2 does not reach the end of file)
           {
                 Read (the column of concept from);
                 Write (the entries corresponding to the row)
                Row2++;
           }
           Remove Conflicts ();
           Detect Cycles ();
           Depth First ();


Row1++;
}

Close (universal_masterthes.csv)
Close (mergestandardlist.csv)
```



	A	B	C	D	E	F
1	conceptFrom	conceptTo	metaOntology	metaName	casosWhat	casosWhy
2	xiaoping	deng_xiaoping	agent		Afghan_Ag	Afghan_Agents_all_1-25-2012 - v6.csv
3	tehrik-italeban	tehrik_i_taleban	agent		CorrectedA	CorrectedAgents_v2.csv
4	tehrik-i-taleban	tehrik_i_taleban	agent		CorrectedA	CorrectedAgents_v2.csv
5	rafsanjani	ali_akbar_hashemi_raf	agent		Afghan_Ag	Afghan_Agents_all_1-25-2012 - v6.csv
6	britton	british	agent		CorrectedA	CorrectedAgents_v2.csv
7	britt	british	agent		CorrectedA	CorrectedAgents_v2.csv
8	bernardo	bernardo_alvarez	agent		Afghan_Ag	Afghan_Agents_all_1-25-2012 - v5.csv
9	beowulf	culture	agent		CorrectedA	CorrectedAgents_v2.csv
10	applebaum	appelbaum	agent		Afghan_Ag	Afghan_Agents_all_1-25-2012 - v5.csv
11	sec-retary	secretary	agent		CorrectedA	CorrectedAgents_v2.csv
12	pashtunestan	pashtun	agent		CorrectedA	CorrectedAgents_v2.csv
13	non-pashtun	pashtun	agent		CorrectedA	CorrectedAgents_v2.csv
14	war-photograph	photographer	agent		CorrectedA	CorrectedAgents_v2.csv
15	ex-service	serviceman	agent		CorrectedA	CorrectedAgents_v2.csv
16	tycoons	tycoon	agent		CorrectedA	CorrectedAgents_v2.csv
17	soldier-politiciar	soldier	agent		CorrectedA	CorrectedAgents_v2.csv
18	scythians	scythian	agent		CorrectedA	CorrectedAgents_v2.csv
19	scottie	scottish	agent		CorrectedA	CorrectedAgents_v2.csv
20	quartermasters	quartermaster	agent		CorrectedA	CorrectedAgents_v2.csv

Fig 24: The domain thesaurus after applying the merge algorithm

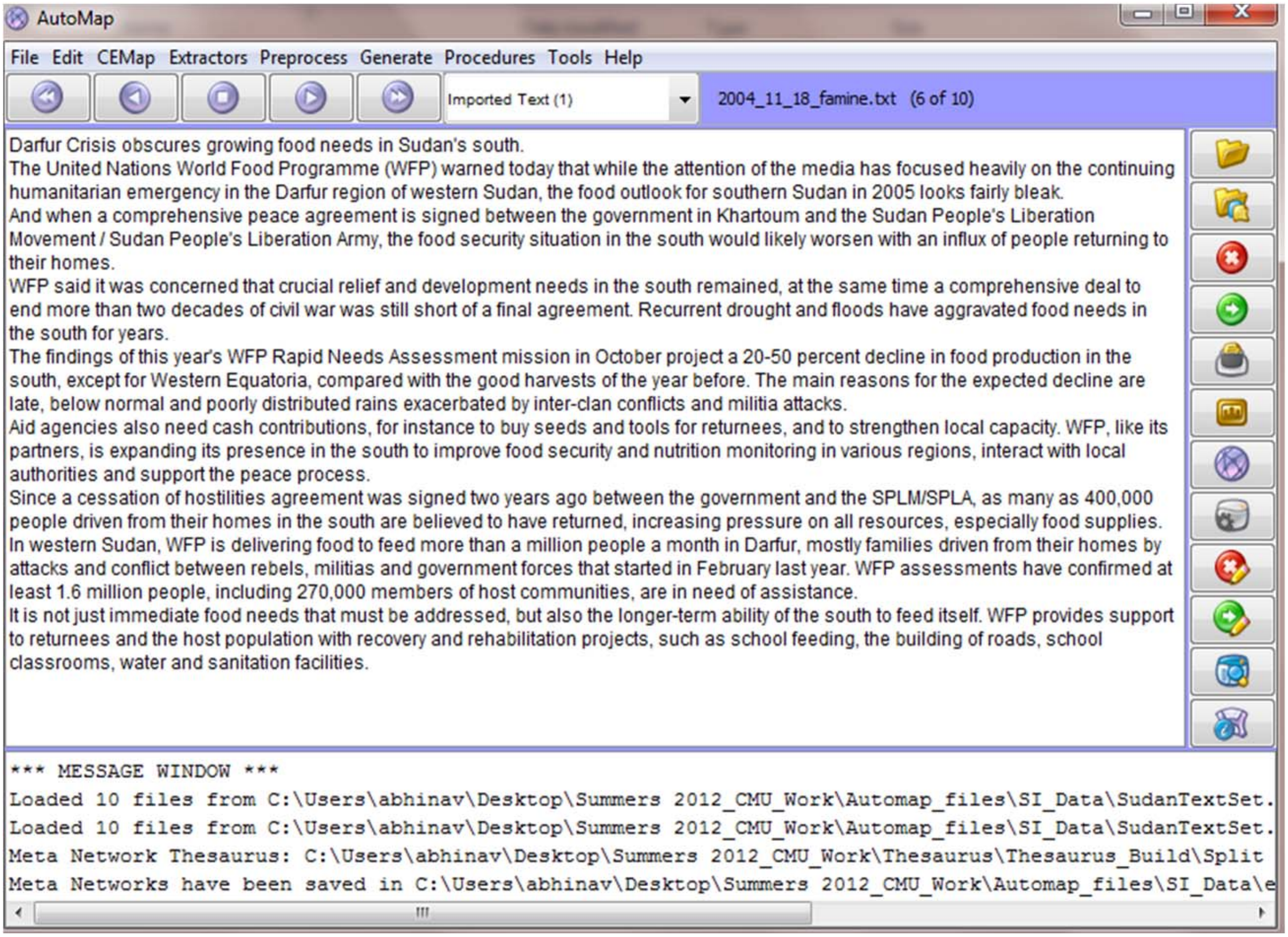


	A	B	C	D	E	F
1	conceptFrom	conceptTo	metaOntology	metaName	casosWhat	casosWhy
2	check these out...	check_these_out...	task		Enron_time	Enron_time_03-03-08.csv
3	business	corporation	delete	generic	nonspecific	nonspecific_gen.csv
4	agricultural	land_water_use	knowledge		universal_n	Entry created because of depth first search f
5	rer_agriculture	land_water_use	knowledge		rer_to_met	rer_to_meta.csv
6	air space	room	location	generic	nonspecific	nonspecific_gen.csv
7	airports airlines	rer_infrastructure	knowledge		universal_n	Entry created because of depth first search f
8	airport	rer_infrastructure	knowledge		universal_n	Entry created because of depth first search f
9	airports	rer_infrastructure	knowledge		universal_n	Entry created because of depth first search f
10	american embassy	us_department_of_	organization		universal_n	Entry created because of depth first search f
11	bird sanctuary	park	location	generic	nonspecific	nonspecific_gen.csv
12	game reserve	park	location	generic	nonspecific	nonspecific_gen.csv
13	wilderness preserve	park	location	generic	nonspecific	nonspecific_gen.csv
14	wildlife preserve	park	location	generic	nonspecific	nonspecific_gen.csv
15	a o r	a_o_r			location_03	location_03-03-2008b.csv
16	area xxx responsibility	a_o_r			location_03	location_03-03-2008b.csv
17	varsity	university	organization	generic	nonspecific	nonspecific_gen.csv
18	bank vault	bank	location	generic	nonspecific	nonspecific_location_kc.csv
19	banks	bank	location	generic	psuira_proj	psuira_master.csv
20	credit union	bank	location	generic	nonspecific	nonspecific_gen.csv

Fig 25: The universal thesaurus after applying the merge algorithm

16 Illustration

After we created a universal thesaurus and the domain thesauri using the above described process, we analyzed the results and hence, the efficiency of this process by running the analysis on a sample data set. For this purpose, we used the Sudan data set.


AutoMap
File Edit CEMap Extractors Preprocess Generate Procedures Tools Help
Imported Text (1)
2004_11_18_famine.txt (6 of 10)

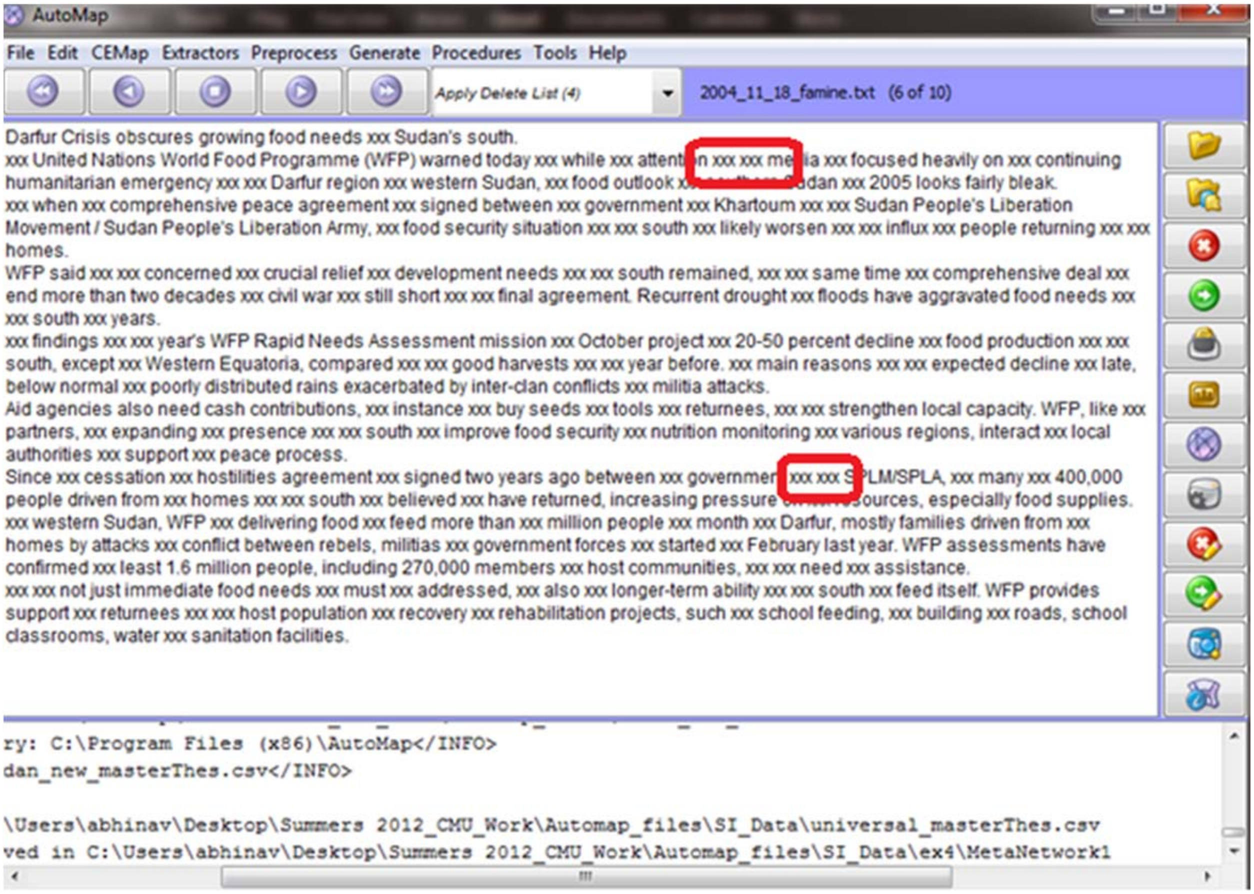
Darfur Crisis obscures growing food needs in Sudan's south.
The United Nations World Food Programme (WFP) warned today that while the attention of the media has focused heavily on the continuing humanitarian emergency in the Darfur region of western Sudan, the food outlook for southern Sudan in 2005 looks fairly bleak.
And when a comprehensive peace agreement is signed between the government in Khartoum and the Sudan People's Liberation Movement / Sudan People's Liberation Army, the food security situation in the south would likely worsen with an influx of people returning to their homes.
WFP said it was concerned that crucial relief and development needs in the south remained, at the same time a comprehensive deal to end more than two decades of civil war was still short of a final agreement. Recurrent drought and floods have aggravated food needs in the south for years.
The findings of this year's WFP Rapid Needs Assessment mission in October project a 20-50 percent decline in food production in the south, except for Western Equatoria, compared with the good harvests of the year before. The main reasons for the expected decline are late, below normal and poorly distributed rains exacerbated by inter-clan conflicts and militia attacks.
Aid agencies also need cash contributions, for instance to buy seeds and tools for returnees, and to strengthen local capacity. WFP, like its partners, is expanding its presence in the south to improve food security and nutrition monitoring in various regions, interact with local authorities and support the peace process.
Since a cessation of hostilities agreement was signed two years ago between the government and the SPLM/SPLA, as many as 400,000 people driven from their homes in the south are believed to have returned, increasing pressure on all resources, especially food supplies.
In western Sudan, WFP is delivering food to feed more than a million people a month in Darfur, mostly families driven from their homes by attacks and conflict between rebels, militias and government forces that started in February last year. WFP assessments have confirmed at least 1.6 million people, including 270,000 members of host communities, are in need of assistance.
It is not just immediate food needs that must be addressed, but also the longer-term ability of the south to feed itself. WFP provides support to returnees and the host population with recovery and rehabilitation projects, such as school feeding, the building of roads, school classrooms, water and sanitation facilities.

```
*** MESSAGE WINDOW ***
Loaded 10 files from C:\Users\abhinav\Desktop\Summers 2012_CMU_Work\Automap_files\SI_Data\SudanTextSet.
Loaded 10 files from C:\Users\abhinav\Desktop\Summers 2012_CMU_Work\Automap_files\SI_Data\SudanTextSet.
Meta Network Thesaurus: C:\Users\abhinav\Desktop\Summers 2012_CMU_Work\Thesaurus\Thesaurus_Build\Split
Meta Networks have been saved in C:\Users\abhinav\Desktop\Summers 2012_CMU_Work\Automap_files\SI_Data\e
```

Fig 26: The Sudan data set.

Firstly, we refine the raw Sudan data by applying the delete lists. The figure shown below is the data set after the universal and the domain delete lists are applied. Since we selected the rhetorical type of deletion; we see **xxx** in place of the deleted words. Note: This is the remainder data after applying both the universal and the Sudan domain delete lists.

Observations: The words which were present either in the universal delete list as noise words or the domain thesaurus as the non-important concepts have been marked as **xxx** in the figure below.


AutoMap
File Edit CEMap Extractors Preprocess Generate Procedures Tools Help
Apply Delete List (4)
2004_11_18_famine.txt (6 of 10)

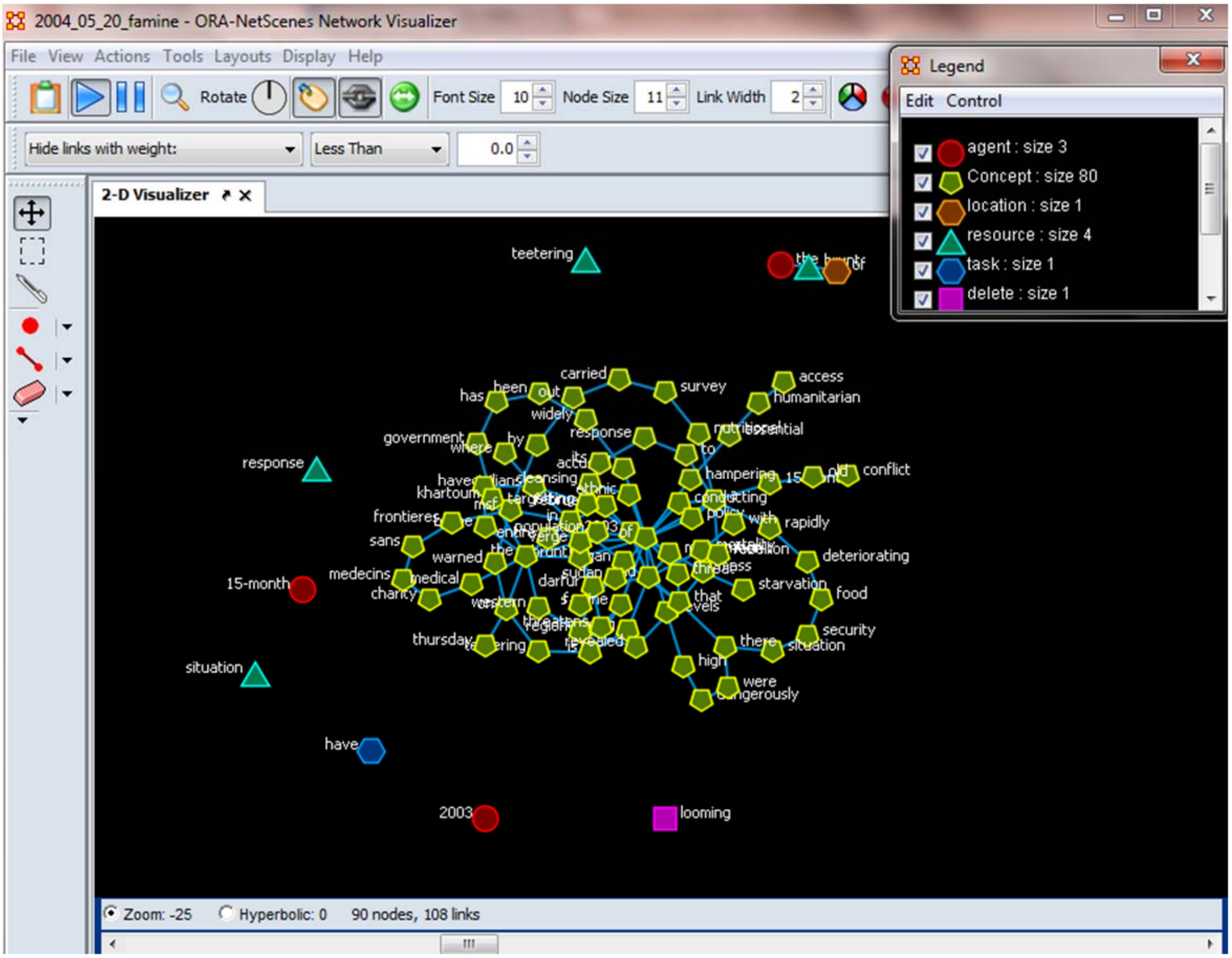
Darfur Crisis obscures growing food needs xxx Sudan's south.
xxx United Nations World Food Programme (WFP) warned today xxx while xxx attention xxx xxx media xxx focused heavily on xxx continuing humanitarian emergency xxx xxx Darfur region xxx western Sudan, xxx food outlook xxx southern Sudan xxx 2005 looks fairly bleak.
xxx when xxx comprehensive peace agreement xxx signed between xxx government xxx Khartoum xxx xxx Sudan People's Liberation Movement / Sudan People's Liberation Army, xxx food security situation xxx xxx south xxx likely worsen xxx xxx influx xxx people returning xxx xxx homes.
WFP said xxx xxx concerned xxx crucial relief xxx development needs xxx xxx south remained, xxx xxx same time xxx comprehensive deal xxx end more than two decades xxx civil war xxx still short xxx xxx final agreement. Recurrent drought xxx floods have aggravated food needs xxx xxx south xxx years.
xxx findings xxx xxx year's WFP Rapid Needs Assessment mission xxx October project xxx 20-50 percent decline xxx food production xxx xxx south, except xxx Western Equatoria, compared xxx xxx good harvests xxx xxx year before. xxx main reasons xxx xxx expected decline xxx late, below normal xxx poorly distributed rains exacerbated by inter-clan conflicts xxx militia attacks.
Aid agencies also need cash contributions, xxx instance xxx buy seeds xxx tools xxx returnees, xxx xxx strengthen local capacity. WFP, like xxx partners, xxx expanding xxx presence xxx xxx south xxx improve food security xxx nutrition monitoring xxx various regions, interact xxx local authorities xxx support xxx peace process.
Since xxx cessation xxx hostilities agreement xxx signed two years ago between xxx government xxx xxx SPLM/SPLA, xxx many xxx 400,000 people driven from xxx homes xxx xxx south xxx believed xxx have returned, increasing pressure on xxx resources, especially food supplies.
xxx western Sudan, WFP xxx delivering food xxx feed more than xxx million people xxx month xxx Darfur, mostly families driven from xxx homes by attacks xxx conflict between rebels, militias xxx government forces xxx started xxx February last year. WFP assessments have confirmed xxx least 1.6 million people, including 270,000 members xxx host communities, xxx xxx need xxx assistance.
xxx xxx not just immediate food needs xxx must xxx addressed, xxx also xxx longer-term ability xxx xxx south xxx feed itself. WFP provides support xxx returnees xxx xxx host population xxx recovery xxx rehabilitation projects, such xxx school feeding, xxx building xxx roads, school classrooms, water xxx sanitation facilities.

```
ry: C:\Program Files (x86)\AutoMap</INFO>
dan_new_masterThes.csv</INFO>

\Users\abhinav\Desktop\Summers 2012_CMU_Work\Automap_files\SI_Data\universal_masterThes.csv
ved in C:\Users\abhinav\Desktop\Summers 2012_CMU_Work\Automap_files\SI_Data\ex4\MetaNetwork1
```

Fig 27: Remainder data after applying the delete lists.

In order to further analyze the data and review the efficiency of our merge process, we created the meta network using the universal thesaurus as the standard thesaurus. The figure shown below shows the meta network when we load the dynetml file in ORA.

Observations: A meta network which consists of some concepts which have been classified according to their meta ontology mappings in the new universal master thesaurus.



2004_05_20_famine - ORA-NetScenes Network Visualizer

Font Size 10 Node Size 11 Link Width 2

Hide links with weight: Less Than 0.0

Legend

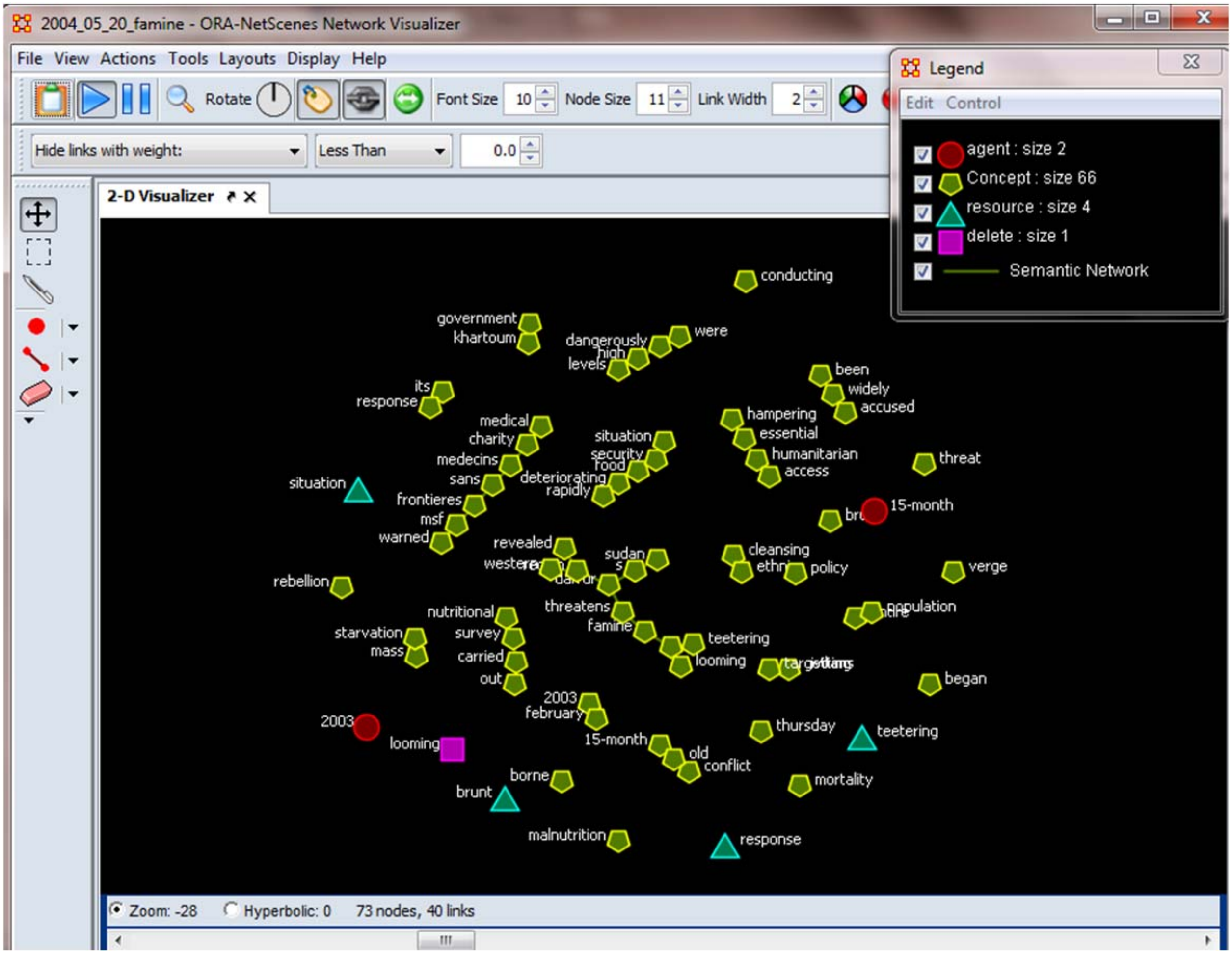
☑ agent : size 3
☑ Concept : size 80
☑ location : size 1
☑ resource : size 4
☑ task : size 1
☑ delete : size 1

Zoom: -25 Hyperbolic: 0 90 nodes, 108 links

Fig 28: Meta network obtained after the universal thesaurus is applied.

Next, we apply the universal delete list for the text refinement and afterwards we create a Dynetml file for the meta network using AutoMap and universal_masterThes.csv as the standard thesaurus.

Observations: The concept size has reduced by around 20% by applying the universal delete list before the Meta network generations. The number of concepts is 66 as compared to the previous observation of 80. Also, we see a reduction in the number of agents by 33% after applying the universal delete list.



2004_05_20_famine - ORA-NetScenes Network Visualizer

Font Size 10 Node Size 11 Link Width 2

Hide links with weight: Less Than 0.0

Legend

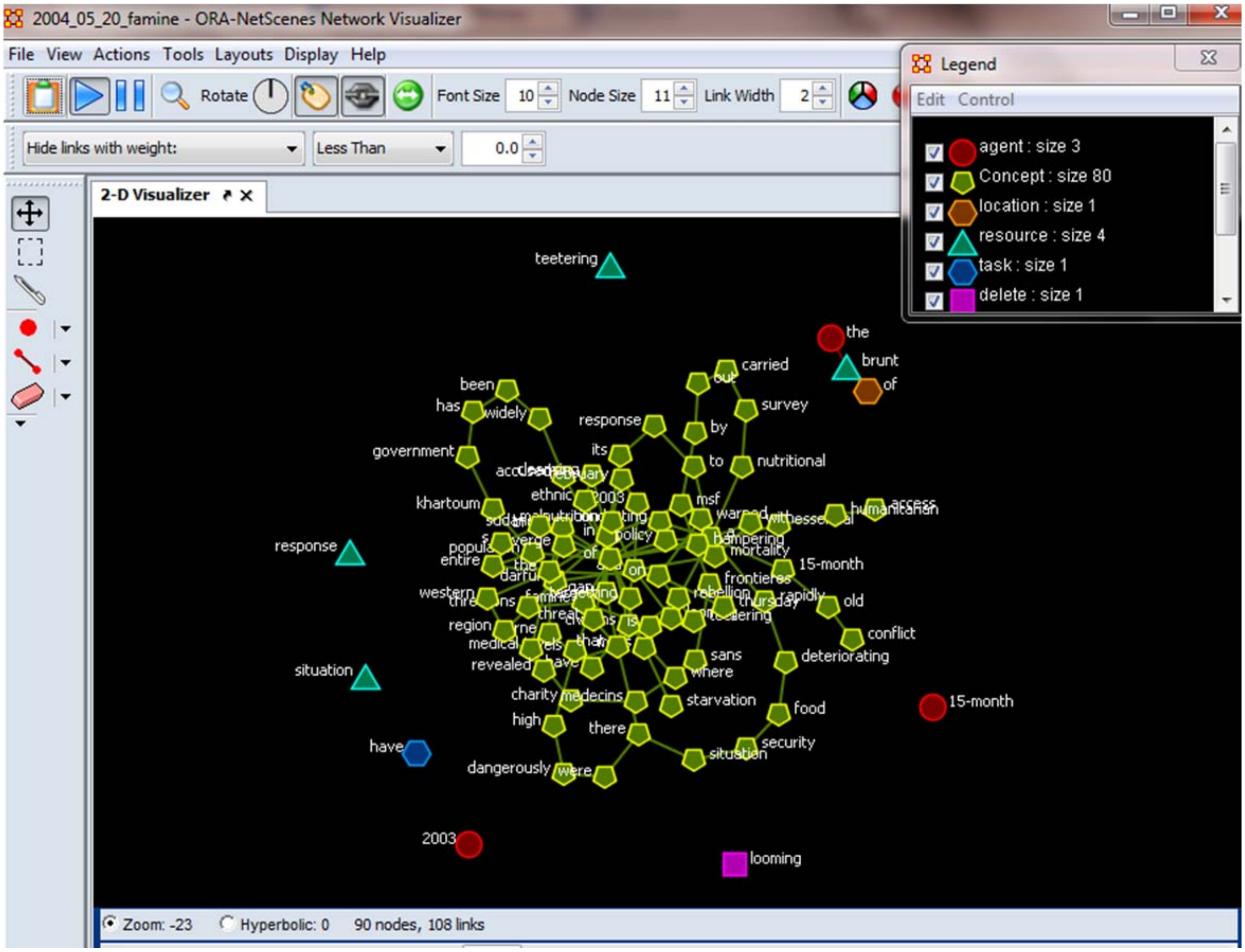
☑ ● agent : size 2
☑ ⬠ Concept : size 66
☑ ▲ resource : size 4
☑ ■ delete : size 1
☑ Semantic Network

○ Zoom: -28 ○ Hyperbolic: 0 73 nodes, 40 links

Fig 29: Meta network obtained after the universal thesaurus is applied along with the universal delete list

The figure below shows the screenshot when the Dynetml file is loaded in ORA. This Dynetml file contains a Meta network which is created when we apply the domain delete list along with the universal_masterThes.csv used as the standard thesaurus.

Observations: A meta network is obtained in which some of the concepts have been classified according to the meta ontologies mapping in the universal master thesaurus. Others are identified as just concepts.



Font Size 10 Node Size 11 Link Width 2

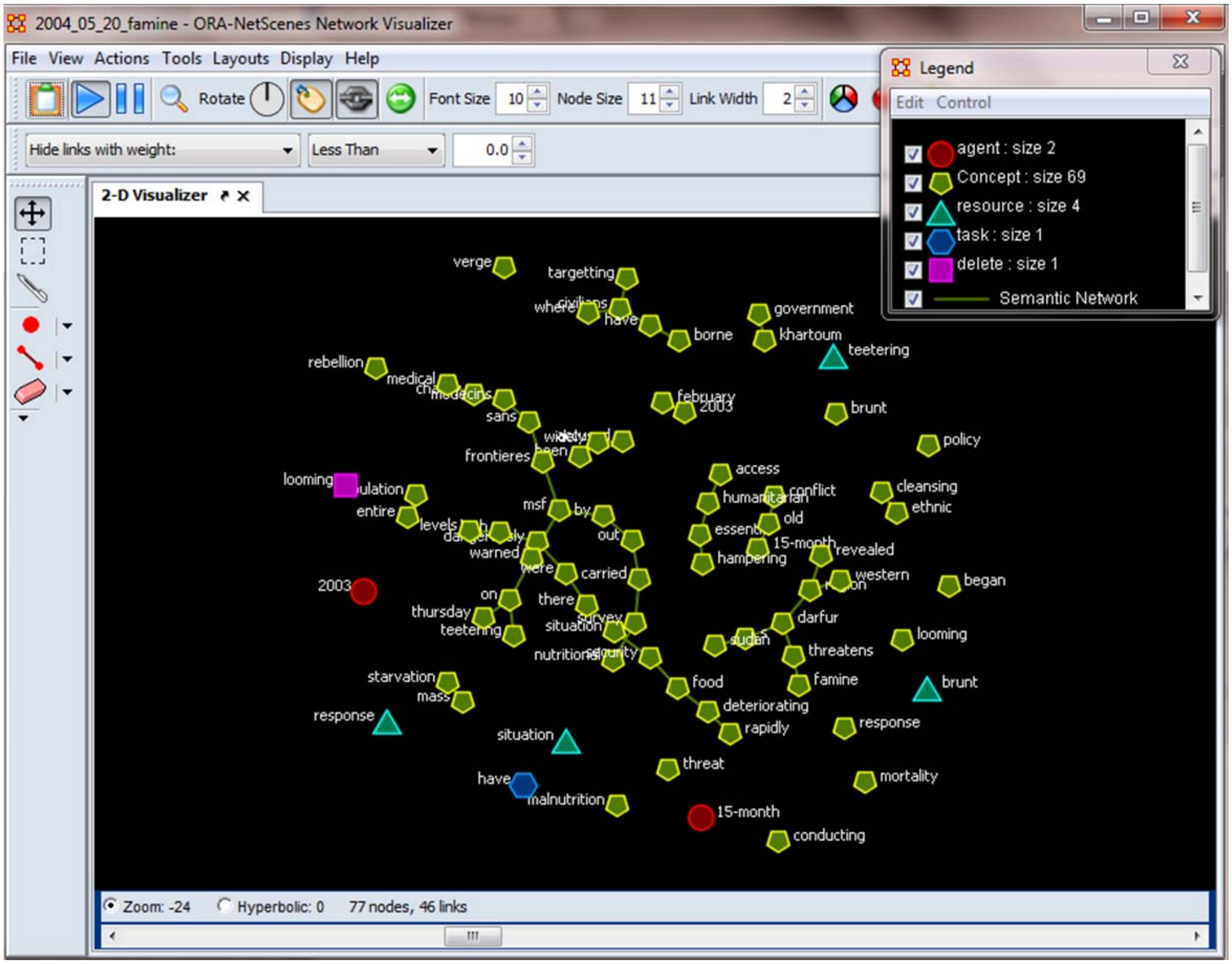
Hide links with weight: Less Than 0.0

☑ agent : size 3
☑ Concept : size 80
☑ location : size 1
☑ resource : size 4
☑ task : size 1
☑ delete : size 1

● Zoom: -23 ○ Hyperbolic: 0 90 nodes, 108 links

Fig 30: Meta network obtained after the domain delete list is applied along with the universal_masterThes.csv used as the standard thesaurus.

Finally, we apply both the universal delete list and the domain delete list. After we have done the text refinement, we apply the domain thesauri as the generalization thesauri. Next, we create a Meta network using the universal_masterThes.csv used as a standard thesaurus.



2004_05_20_famine - ORA-NetScenes Network Visualizer

Font Size 10 Node Size 11 Link Width 2

Hide links with weight: Less Than 0.0

Legend

☑ ● agent : size 2
☑ ⬟ Concept : size 69
☑ ▲ resource : size 4
☑ ⬢ task : size 1
☑ ■ delete : size 1
☑ — Semantic Network

Zoom: -24 Hyperbolic: 0 77 nodes, 46 links

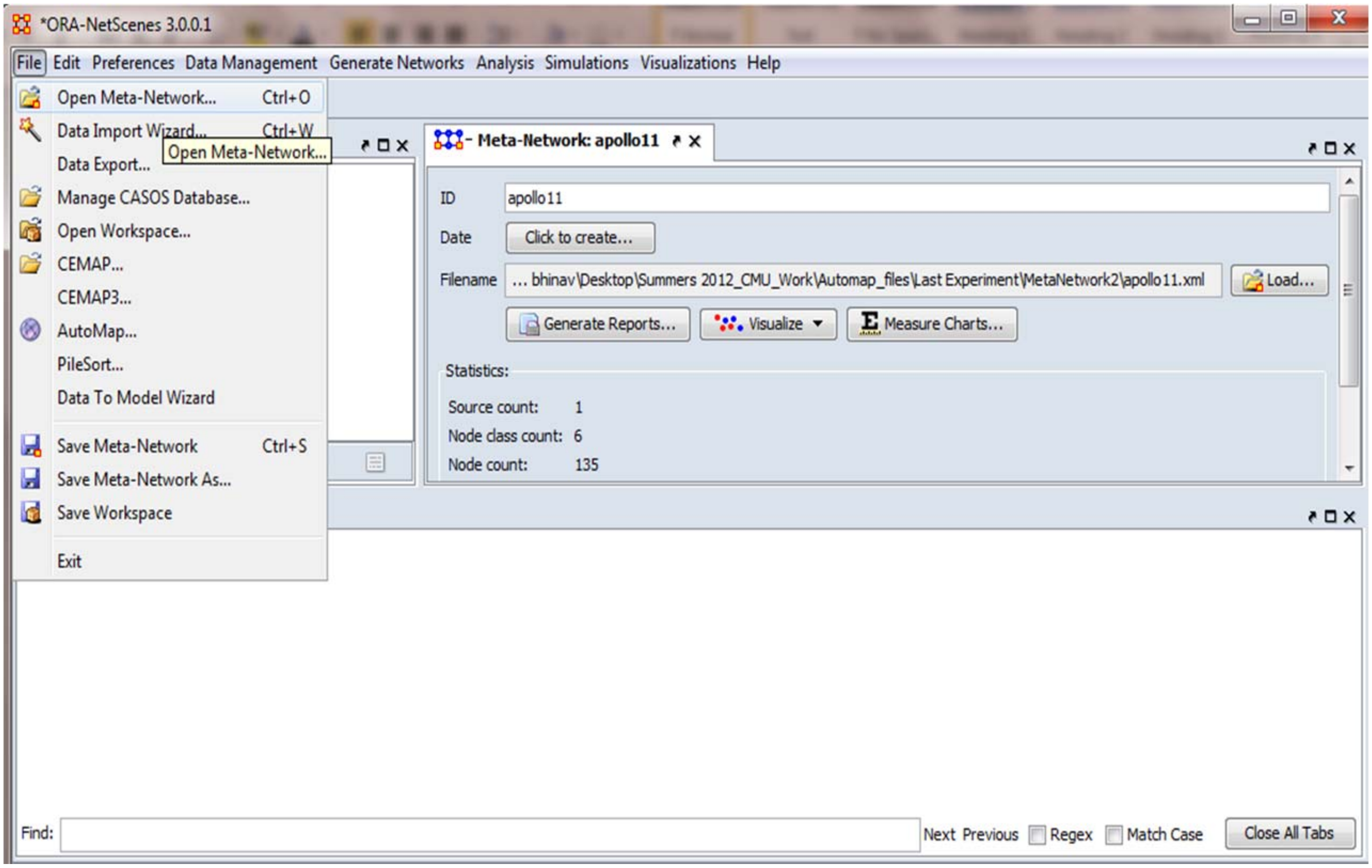
Fig 31: Meta network obtained by using both the universal and the domain delete lists, and using the domain thesauri and the universal thesaurus as the standard thesaurus.

17 Creating and Applying Change Files in ORA

In this section, we discuss how to create the change files in ORA and apply the changes in ORA itself. The users can use this approach to modify the Meta networks extracted from the texts using AutoMap. By using this approach, the users can directly apply the changes to Meta network without using AutoMap for modifying the nodes, links and the networks.

17.1 Creating Change files in ORA

Step 1: Load the Meta network in ORA. The users can go to **file->Open Meta-network** and then browse to the location of the required network files or just simply drag and drop the dynetml file



ID: apollo11

Date: Click to create...

Filename: ... bhinav\Desktop\Summers 2012_CMU_Work\Automap_files\Last Experiment\MetaNetwork2\apollo11.xml

Statistics:

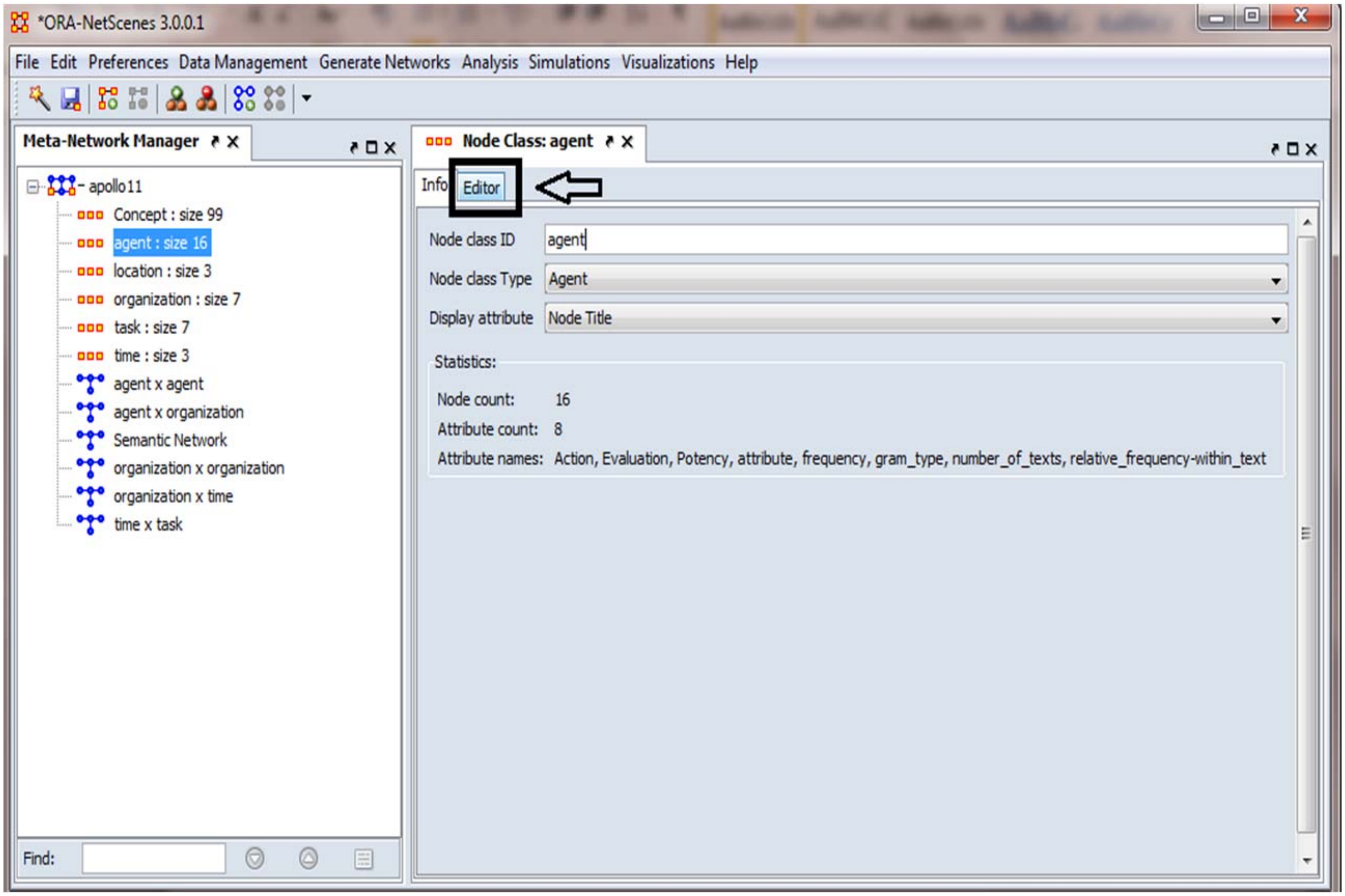
Source count: 1

Node class count: 6

Node count: 135

Fig 32: Opening Meta network in ORA.

Step 2: Once the Meta network is opened in ORA, select the node class in which you want to make any changes. If the node class is uncertain to the users, they can simply just click on the node class “**concepts”**, which includes the concepts from all the node classes. After this, the user should go to the **“editor”** tab.

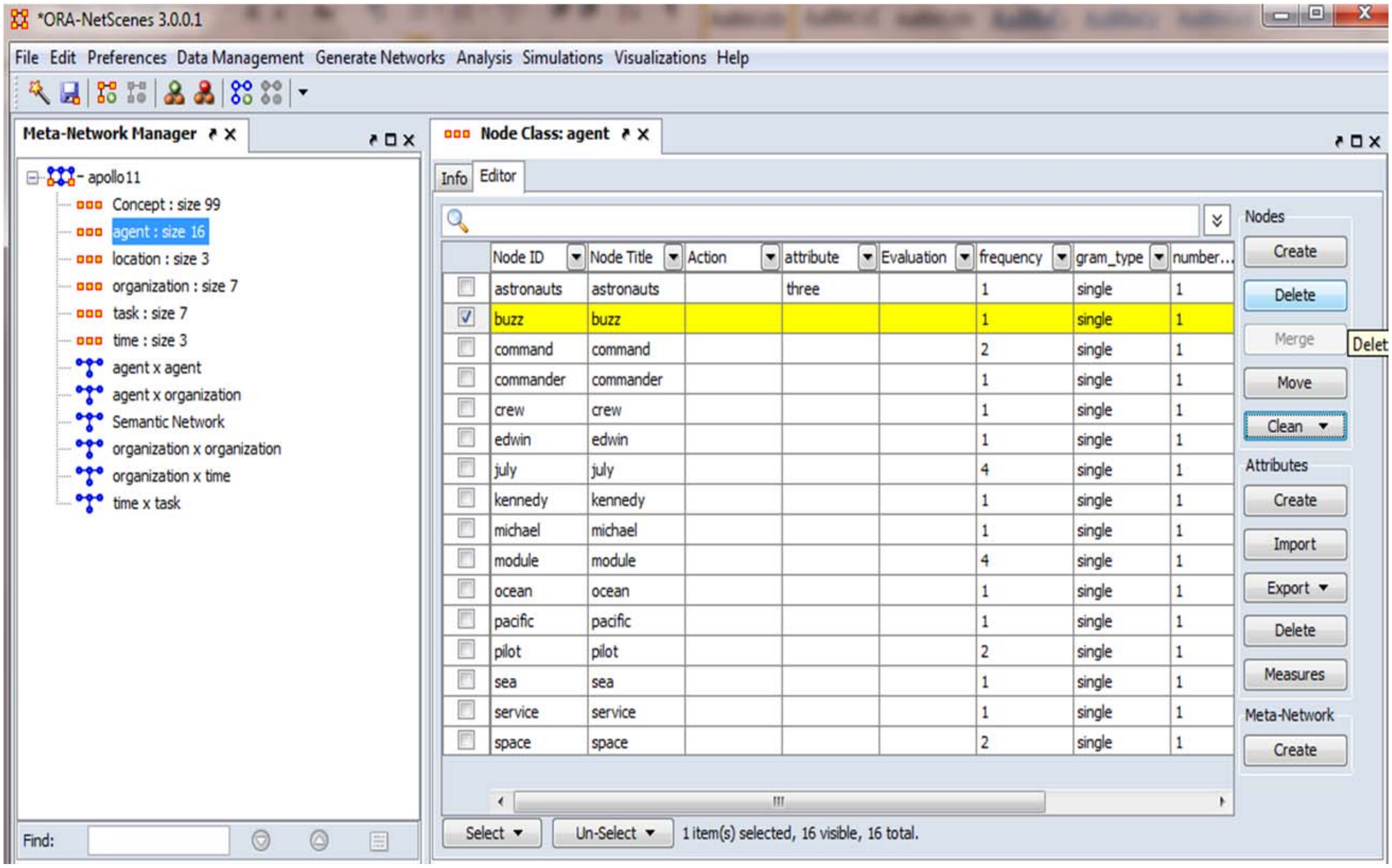


Node class ID: agent
Node class Type: Agent
Display attribute: Node Title

Statistics:
Node count: 16
Attribute count: 8
Attribute names: Action, Evaluation, Potency, attribute, frequency, gram_type, number_of_texts, relative_frequency-within_text

Fig 33: Modifying the nodes in ORA.

Step 3: Once the editor tab is opened corresponding to the node class where the required node class changes are to be made, the user should select the node on which the changes are to made. When the node is selected, a check appears on the left and the row is highlighted in yellow. After this, the users can select from a number of node class operations. These operations are Delete node, moving the node to a different node class, copying the node to a different node class etc. For example: we want to delete the agent **“buzz”**. We select the node buzz and delete this node from the network.



	Node ID	Node Title	Action	attribute	Evaluation	frequency	gram_type	number...
□	astronauts	astronauts		three		1	single	1
☑	buzz	buzz				1	single	1
□	command	command				2	single	1
□	commander	commander				1	single	1
□	crew	crew				1	single	1
□	edwin	edwin				1	single	1
□	july	july				4	single	1
□	kennedy	kennedy				1	single	1
□	michael	michael				1	single	1
□	module	module				4	single	1
□	ocean	ocean				1	single	1
□	pacific	pacific				1	single	1
□	pilot	pilot				2	single	1
□	sea	sea				1	single	1
□	service	service				1	single	1
□	space	space				2	single	1

1 item(s) selected, 16 visible, 16 total.

Fig 34: Deleting a node in ORA.

Step 4: After the nodes are modified in the Meta network, we are ready to save this change list. Saving this change list helps us to use this list for future reference and modifying network afterwards. The users should go to **Data Management-> Meta network clean**. A window pops out and then the user should select the required Meta network wherein the node changes were made.

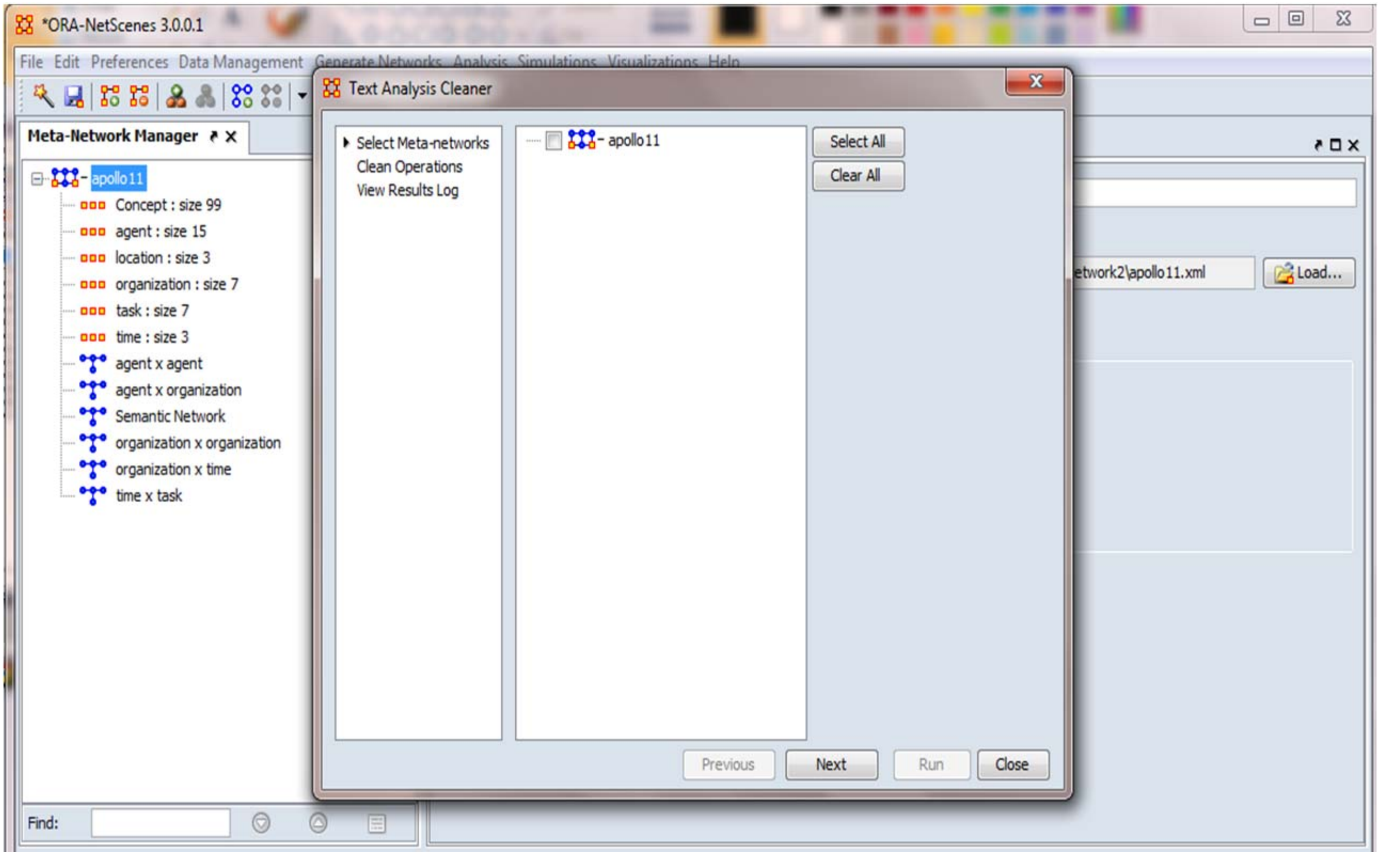
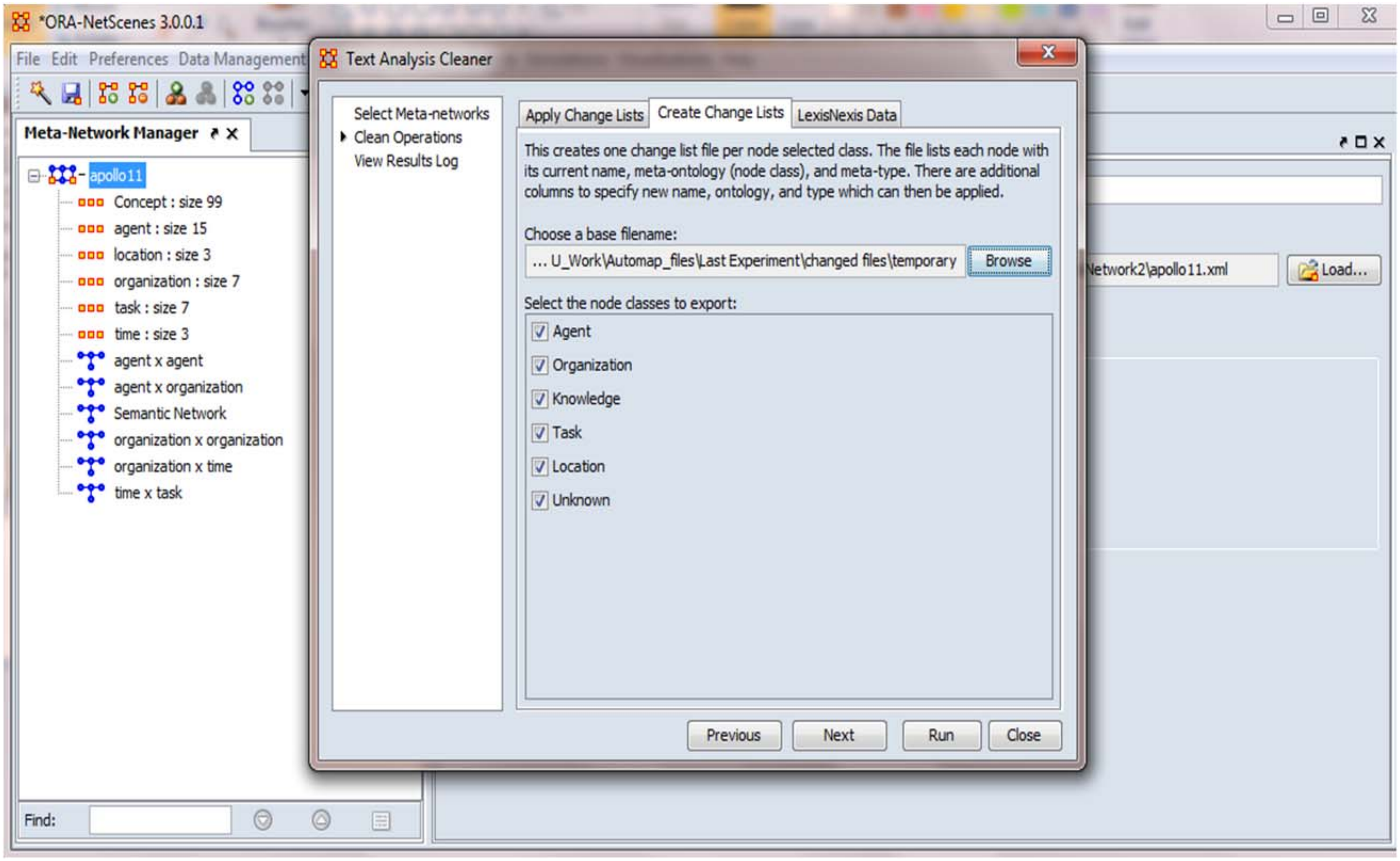


Fig 35: Creating a change list in ORA.

Step 5: Select the Meta network and then click next. In the clean operations menu, select the tab **create change list**. The users need to provide a base name for the changed file and then select the node classes for which the changed file is required. Then click **RUN**. After the process is completed, click **Close**.



This creates one change list file per node selected class. The file lists each node with its current name, meta-ontology (node class), and meta-type. There are additional columns to specify new name, ontology, and type which can then be applied.

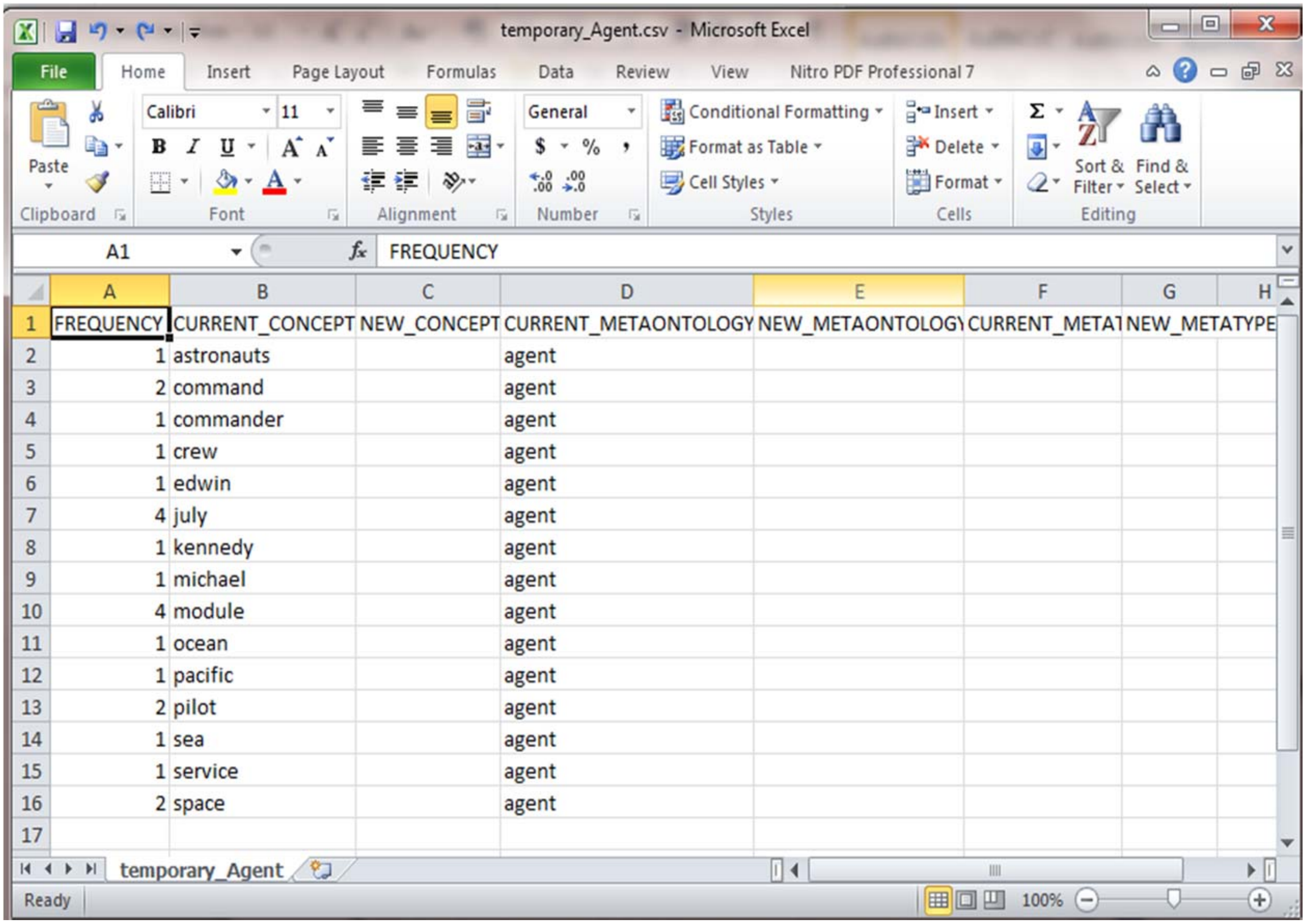
Choose a base filename: ... U_Work\Automap_files\Last Experiment\changed files\temporary

Select the node classes to export:

- ☑ Agent
- ☑ Organization
- ☑ Knowledge
- ☑ Task
- ☑ Location
- ☑ Unknown

Fig 36: Saving a change list in ORA.

Step 6: Browse to the location of the change files and explore the change files. We see that the change files resemble the reduced format (See section 3.5 on page).



A1 f_x FREQUENCY

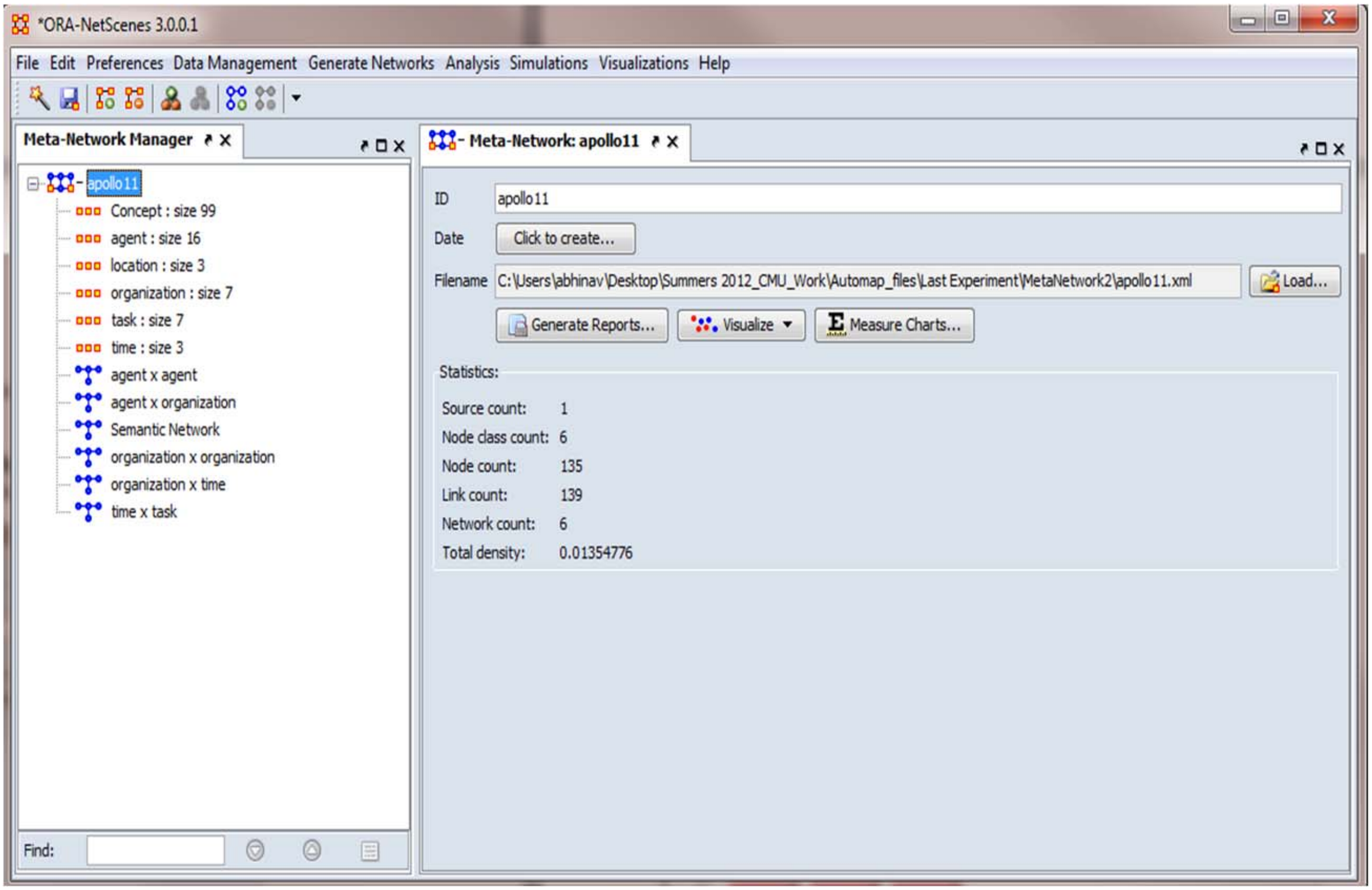
	A	B	C	D	E	F	G
1	FREQUENCY	CURRENT_CONCEPT	NEW_CONCEPT	CURRENT_METAONTOLOGY	NEW_METAONTOLOG	CURRENT_META	NEW_METATYPE
2	1	astronauts		agent			
3	2	command		agent			
4	1	commander		agent			
5	1	crew		agent			
6	1	edwin		agent			
7	4	july		agent			
8	1	kennedy		agent			
9	1	michael		agent			
10	4	module		agent			
11	1	ocean		agent			
12	1	pacific		agent			
13	2	pilot		agent			
14	1	sea		agent			
15	1	service		agent			
16	2	space		agent			
17							

Fig 37: The format for the change file.

17.2 Applying change files in ORA

We can also apply the change files in ORA to update the Meta networks. We will describe a step-by-step approach to apply the change files in ORA.

Step 1: Load the Meta network in ORA. The users can go to **file->Open Meta-network** and then browse to the location of the required network files or just simply drag and drop the dynetml file



apollo11
- Concept : size 99
- agent : size 16
- location : size 3
- organization : size 7
- task : size 7
- time : size 3
- agent x agent
- agent x organization
- Semantic Network
- organization x organization
- organization x time
- time x task

ID: apollo11

Date: Click to create...

Filename: C:\Users\abhinav\Desktop\Summers 2012_CMU_Work\Automap_files\Last Experiment\MetaNetwork2\apollo11.xml

Statistics:

Source count:	1
Node class count:	6
Node count:	135
Link count:	139
Network count:	6
Total density:	0.01354776

Fig 38: Meta network opened in ORA

Step 2: The users should go to **Data Management-> Meta network clean**. A window pops out and then the user should select the required Meta network wherein the node changes are to be made.

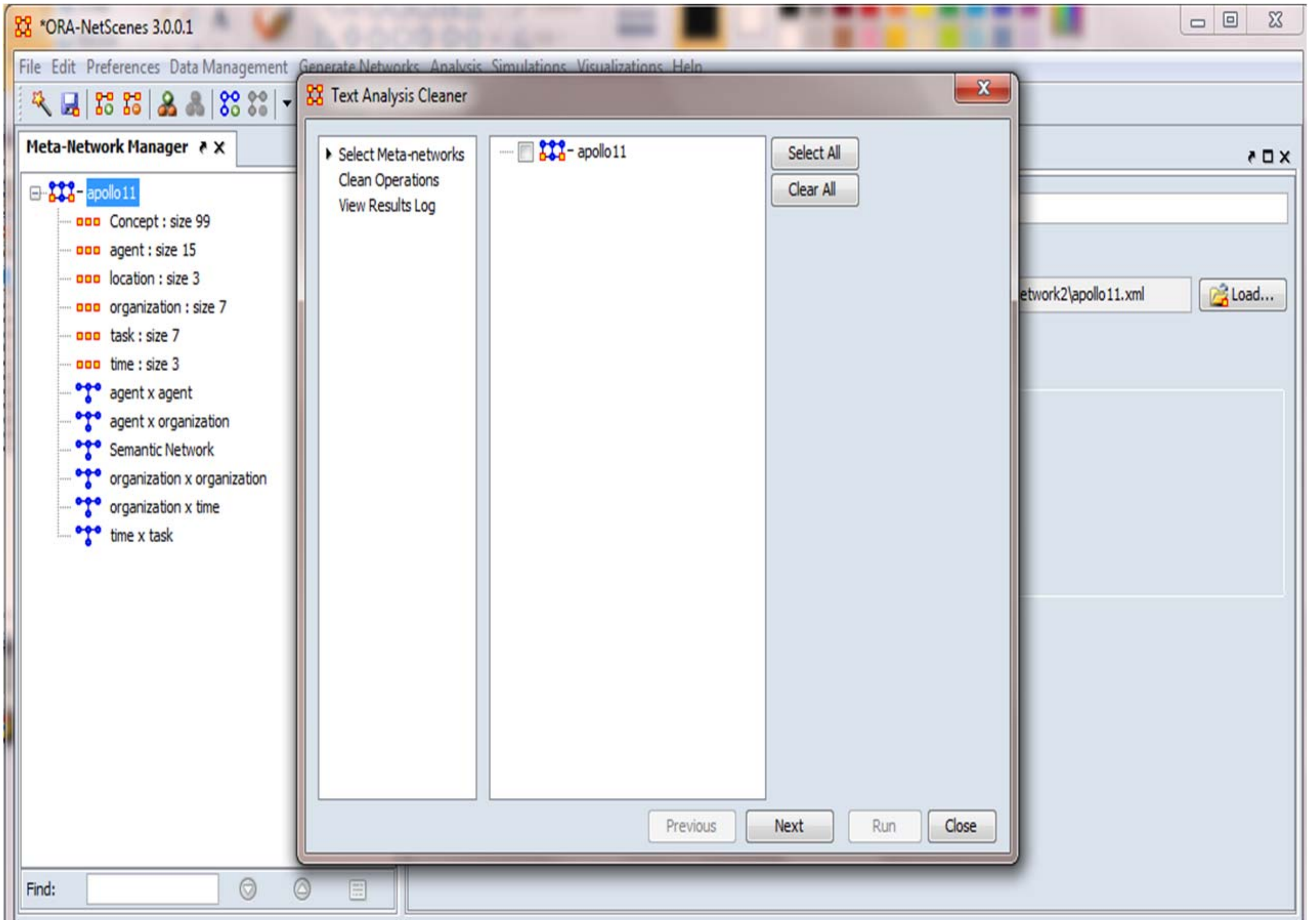
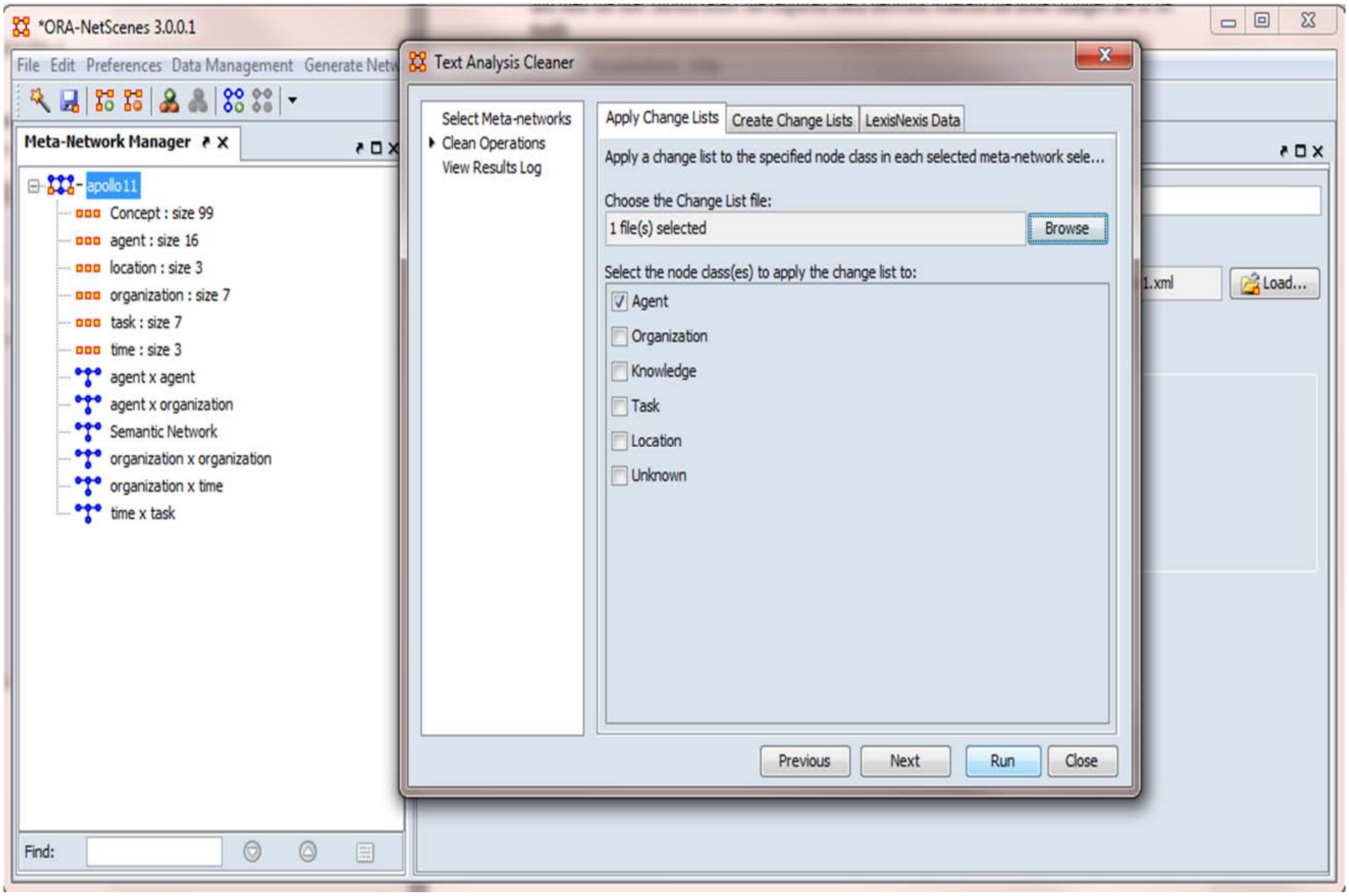


Fig 39: Selecting the Meta network where the changes are to be applied.

Step 3: Select the Meta network and then click next. In the clean operations menu, select the tab **Apply change list**. The users need to provide the location for the changed list file and then select all the node classes for which the changed file in which the changes are required. Then click **RUN**. After the process is completed, click **Close**.



Apply a change list to the specified node class in each selected meta-network sele...

Choose the Change List file: 1 file(s) selected

Select the node class(es) to apply the change list to:

☑ Agent
☐ Organization
☐ Knowledge
☐ Task
☐ Location
☐ Unknown

Fig 40: Applying the change list in ORA.

18 Results

After we successfully created a universal thesaurus and several other domain thesauri for various projects (such as Enron, Catnet, Sudan etc.), in order to analyze the efficiency of the merge and split process, we compared the results of the Catnet data set using the standard thesaurus (i.e., the old version of the universal merged thesaurus) and the new Universal merged thesaurus created with the process described in this report.

The data set we used to compare the results of the D2M process [3] using the standard thesaurus and the merged thesaurus were related to the interviews that took place with one of the members of the Al-mujaharoon group. To get a better idea of the difference of the results between the outputs of the D2M run [3] using the standard and the merged thesaurus, we generated the key entity reports [2] for the two D2M [3] outputs.

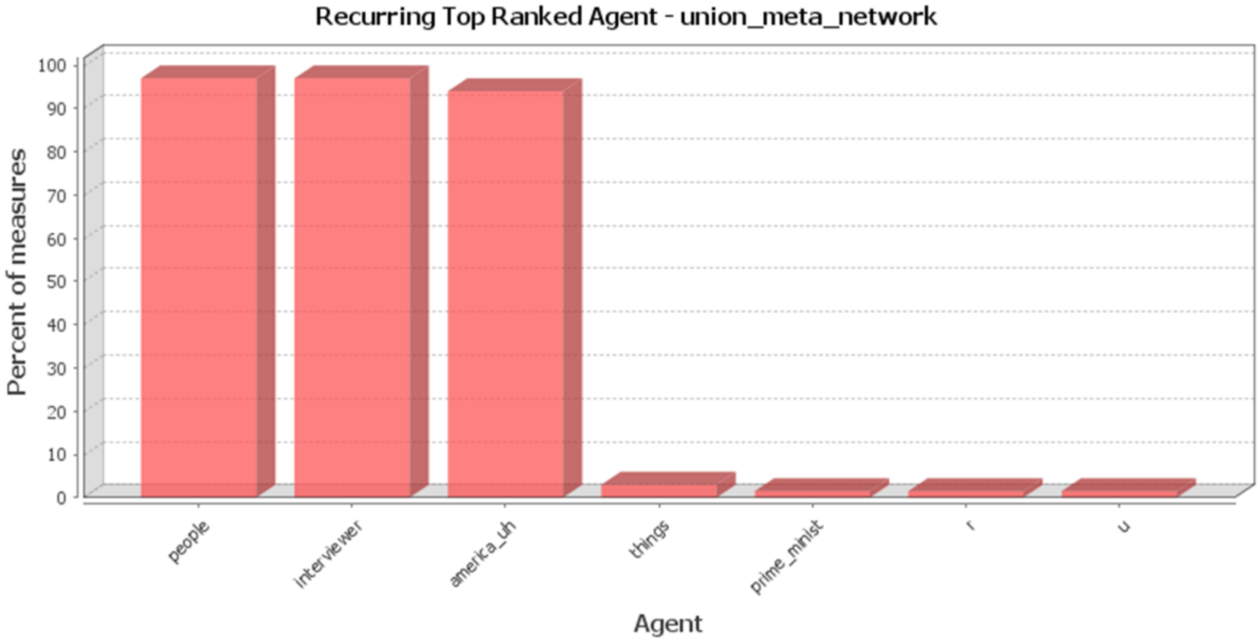


Fig 41: Top ranked agents using the Standard thesaurus

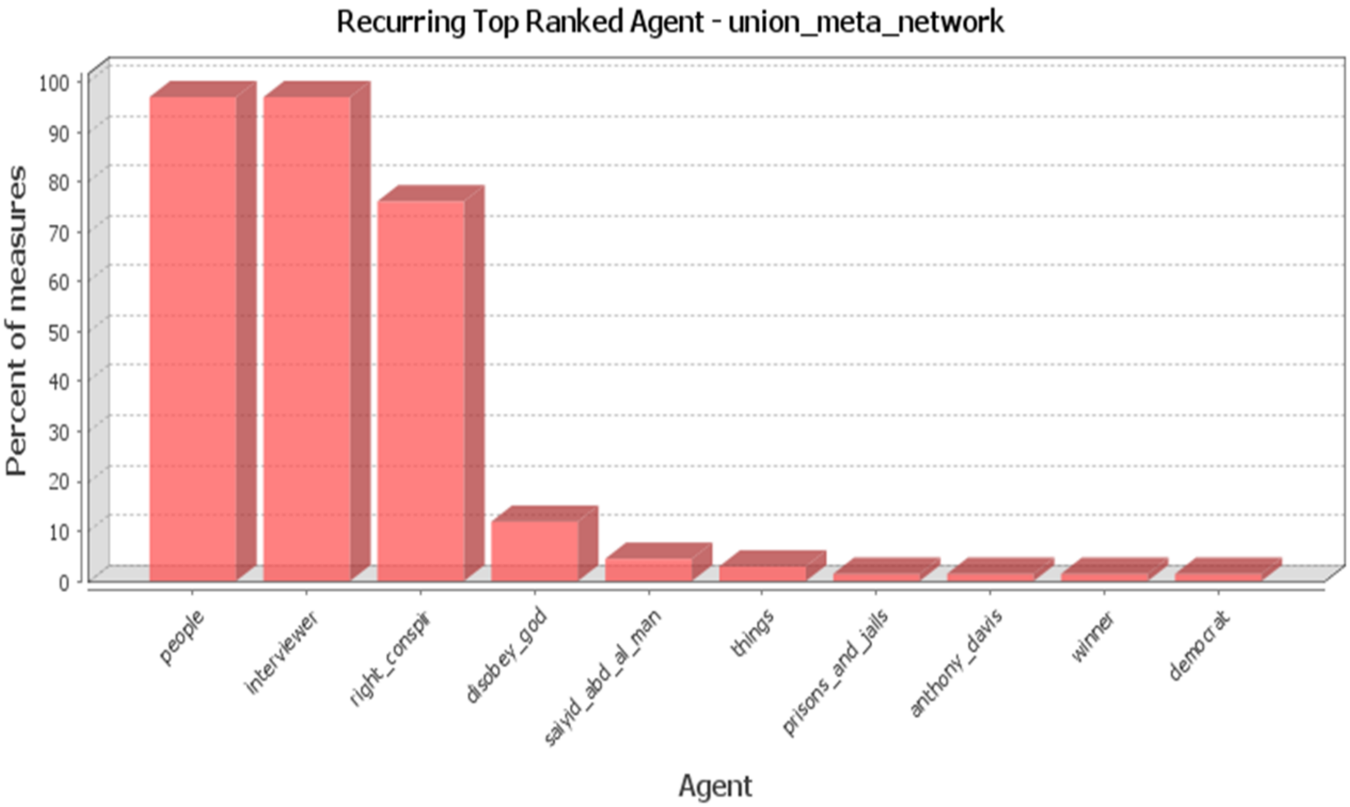


Fig 42: Top ranked agents using the Merged Thesaurus

The above two figures depict the recurring top ranked agents for the D2M run using the standard thesauri and the D2M run using the merged thesauri. The bar graph depicts the top 10 recurring agents in the 24 measures used for a qualitative analysis in the key report.

The 24 measures are as follows:

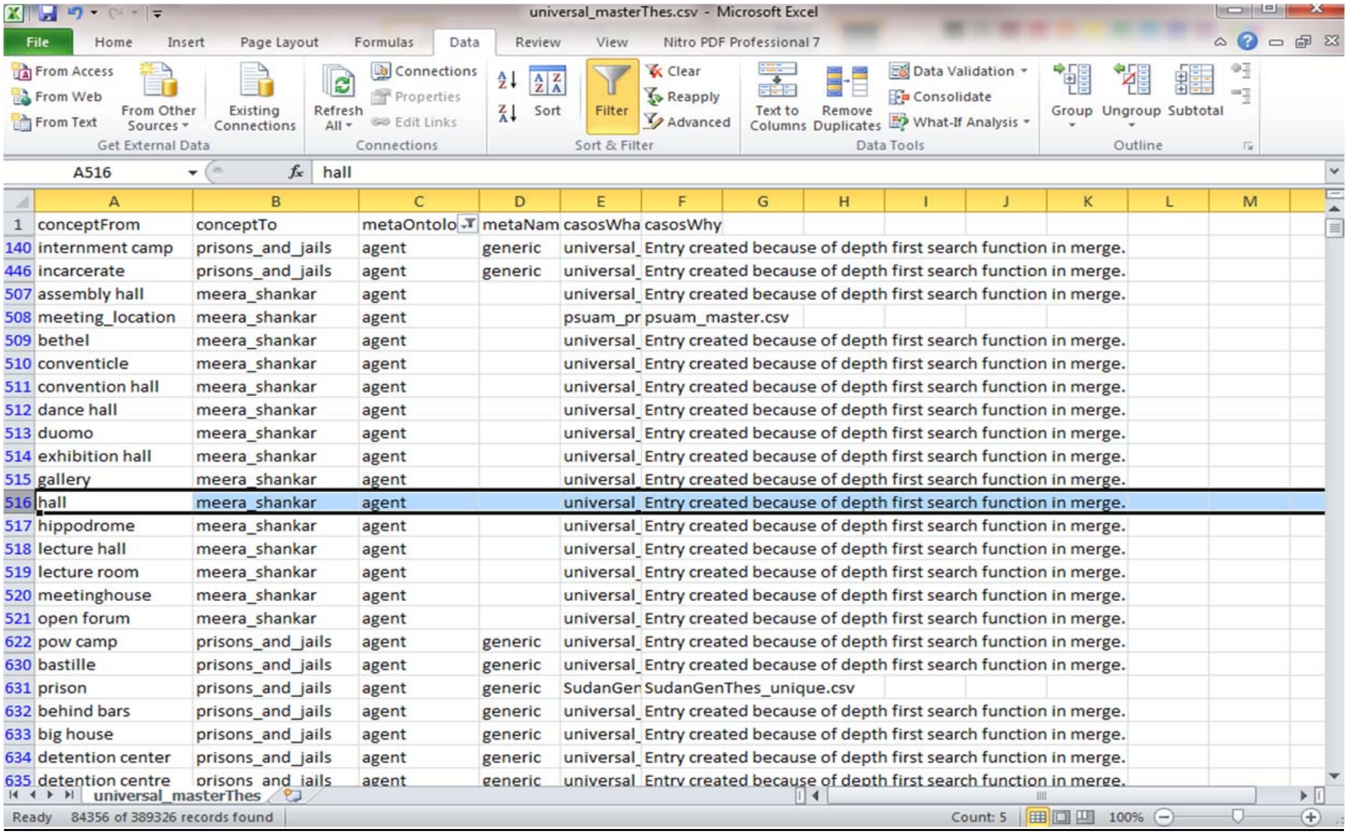
1. Emergent Leader (cognitive demand)
2. In-the-Know (total degree centrality)
3. Number of Cliques (clique count)
4. Most Knowledge (row degree centrality)
5. Most Resources (row degree centrality)
6. Leader of Strong Clique (eigenvector centrality)
7. Leader of Strong Clique Per Component (eigenvector centrality per component)
8. Acts as a Hub (hub centrality)
9. Acts as an Authority (authority centrality)
10. Potentially Influential (between ness centrality)
11. Potentially Influential Links (edge between ness centrality)
12. Connects Groups (high between ness and low degree)
13. Specialization - knowledge (relatively unique)
14. Complete Exclusivity - knowledge (complete exclusivity)
15. Specialization - resource (relatively unique)

16. Complete Exclusivity - resource (complete exclusivity)
17. Specialization - task (relatively unique)
18. Complete Exclusivity - task (complete exclusivity)
19. Specialization - event (relatively unique)
20. Complete Exclusivity - event (complete exclusivity)
21. Specialization - location (relatively unique)
22. Complete Exclusivity - location (complete exclusivity)
23. Workload (actual based on knowledge and resource)
24. Group Awareness (shared situation awareness)

It was observed that the top 10 recurring agents in the standard D2M run included agents, **r** and **u** which were completely irrelevant when considering the key agents. Items such as these indicate errors in the standard thesauri.

In contrast, the observations with regard to the top 10 key agents emergent in the key entity report using the merged thesaurus run were more relevant and informative. The new key entity report was able to do away with erroneous entities in the top 10 agents like “r” and “u”.

Secondly, agents like “saiyid_abd_al_man” and ”Anthony Davis” which emerged as two of the top 10 ranked agents using the merged thesaurus were not even present in the key entity report generated using the standard thesaurus.



A516 fx hall

	A	B	C	D	E	F
1	conceptFrom	conceptTo	metaOntolo	metaNam	casosWha	casosWhy
140	internment camp	prisons_and_jails	agent	generic	universal_	Entry created because of depth first search function in merge.
446	incarcerate	prisons_and_jails	agent	generic	universal_	Entry created because of depth first search function in merge.
507	assembly hall	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
508	meeting_location	meera_shankar	agent		psuam_pr	psuam_master.csv
509	bethel	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
510	conventicle	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
511	convention hall	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
512	dance hall	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
513	duomo	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
514	exhibition hall	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
515	gallery	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
516	hall	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
517	hippodrome	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
518	lecture hall	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
519	lecture room	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
520	meetinghouse	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
521	open forum	meera_shankar	agent		universal_	Entry created because of depth first search function in merge.
622	pow camp	prisons_and_jails	agent	generic	universal_	Entry created because of depth first search function in merge.
630	bastille	prisons_and_jails	agent	generic	universal_	Entry created because of depth first search function in merge.
631	prison	prisons_and_jails	agent	generic	SudanGen	SudanGenThes_unique.csv
632	behind bars	prisons_and_jails	agent	generic	universal_	Entry created because of depth first search function in merge.
633	big house	prisons_and_jails	agent	generic	universal_	Entry created because of depth first search function in merge.
634	detention center	prisons_and_jails	agent	generic	universal_	Entry created because of depth first search function in merge.
635	detention centre	prisons_and_jails	agent	generic	universal_	Entry created because of depth first search function in merge.

Ready 84356 of 389326 records found

Fig 43: Universal_masterThes.csv

19 Future Directions

In the near future, we would want this process to get more automated so as to reduce the human effort involved in combining all the files in a single directory and manually listing the files in the splitcontrol.csv and mergecontrol.csv.

One of the major problems which exist in the existing process is that the files in the sub directory will not be read from the input directory. This means that if we want to keep some files separate and to merge as well, this would be difficult and cumbersome.

The existing process takes the input files in a particular format, i.e. the master thesaurus format. If the user provides the merge routine with a file in any other format, this file would not get merged. So, one of the improvements could be to automatically convert all the input files into the required format and then do the merge step.

As stated, the merge and the split process takes the files as input in master thesaurus format which is less informative then the reduced format type (which is a 7 column type thesaurus). So, another improvement with regard to the formats would be to support input files in the change format so that the output in turn becomes more informative and much more descriptive.

We would also like to use the change files created from ORA directly in AutoMap, so that gives the users a method to manipulate the changes in AutoMap and incorporate them in ORA and vice-versa.

Additional drawbacks of this merge and split process involve the use of depth first search in the merge procedure. This means that if a concept like hall is general it might get linked to something like meera_shankar because of the function depth first search. This in turn might create some shocking and unexpected results in the key entity reports [2] and hence, the overall analysis.

20 References

[1] Carley, Kathleen & Columbus, Dave & Azoulay, Ariel (2012). *AutoMap User's Guide 2012* Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report CMU-ISR-12-106.

[2] Carley, Kathleen & Columbus, Dave (2012). *Basic Lessons in ORA and AutoMap 2012* Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report CMU-ISR-12-107.

[3] Carley, Kathleen & Bigrigg, Michael & Diallo, Bouba. *Data-to-Model: A Mixed Initiative Approach for Rapid Ethnographic Assessment*. Carnegie Mellon University, School of Computer Science, Institute for Software Research, Technical Report.