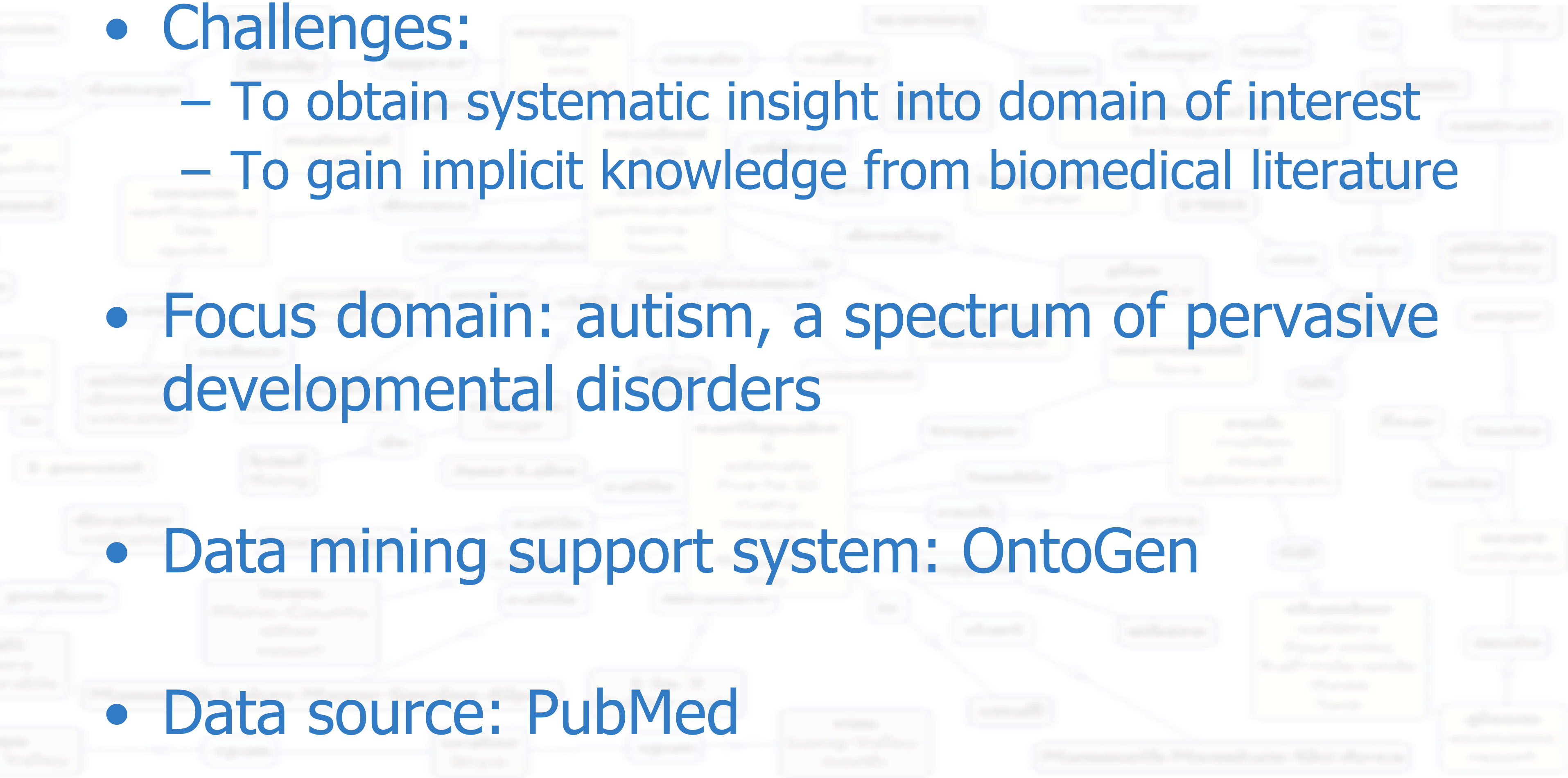
DATA MINING BY ONTOLOGIES CONSTRUCTION IN BIOMEDICAL DOMAIN

Ingrid Petrič
Tanja Urbančič
Bojan Cestnik

DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Overview



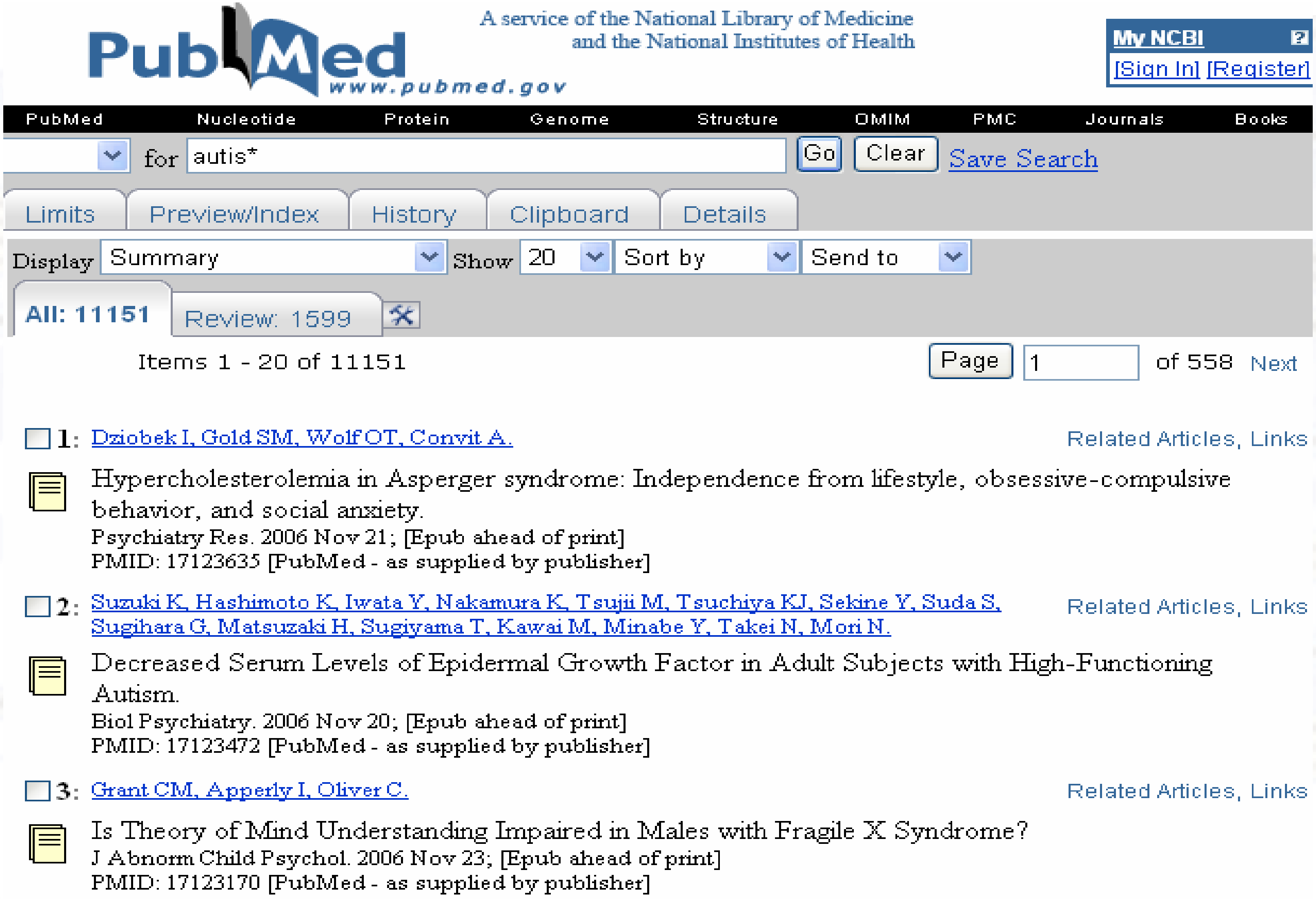
- Challenges:
  - To obtain systematic insight into domain of interest
  - To gain implicit knowledge from biomedical literature

- Focus domain: autism, a spectrum of pervasive developmental disorders

- Data mining support system: OntoGen

- Data source: PubMed

Input data source



Items 1 - 20 of 11151

□ 1: Dziobek I, Gold SM, Wolf OT, Convit A. Related Articles, Links

Hypercholesterolemia in Asperger syndrome: Independence from lifestyle, obsessive-compulsive behavior, and social anxiety.
Psychiatry Res. 2006 Nov 21; [Epub ahead of print]
PMID: 17123635 [PubMed - as supplied by publisher]

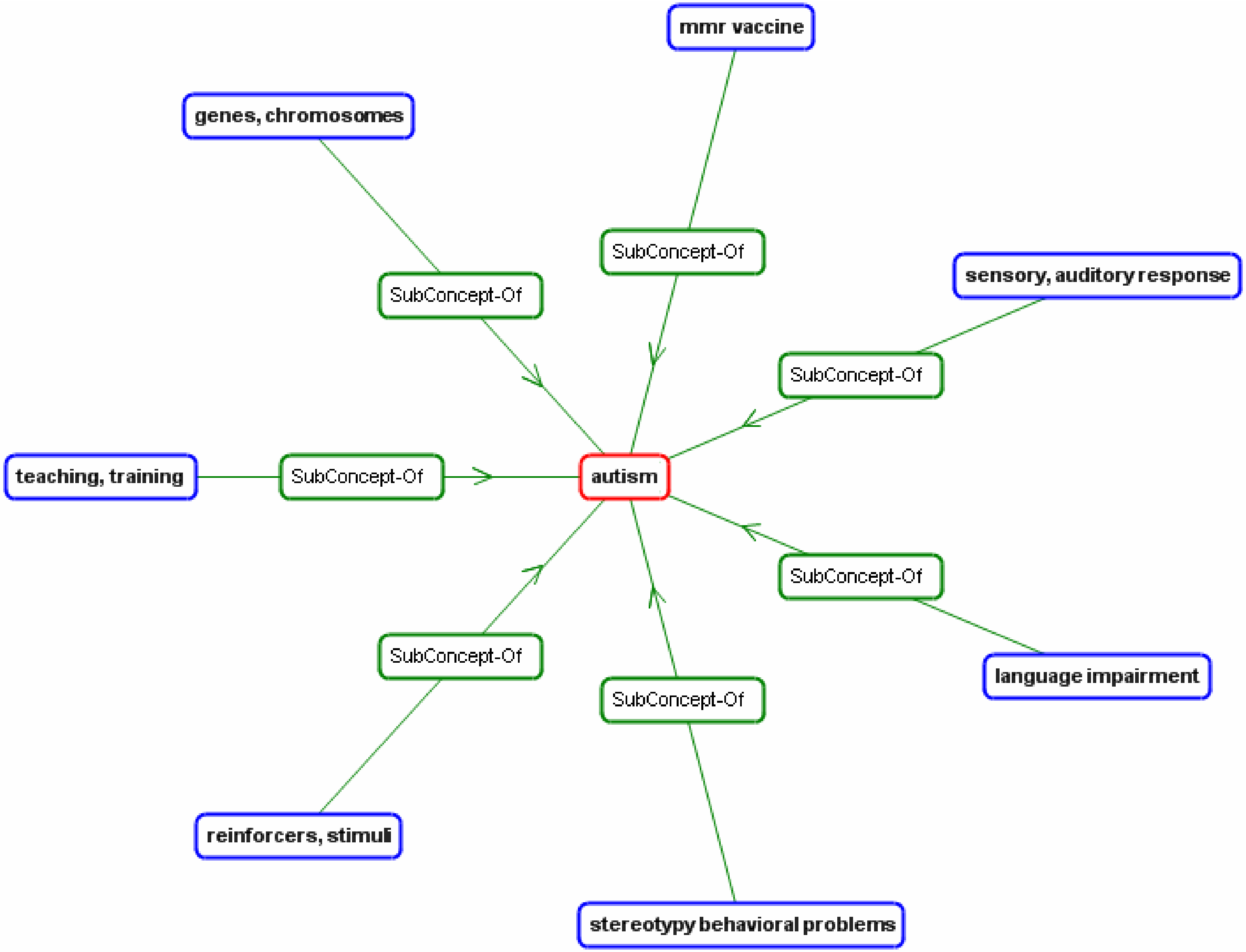
□ 2: Suzuki K, Hashimoto K, Iwata Y, Nakamura K, Tsujii M, Tsuchiya KJ, Sekine Y, Suda S, Sugihara G, Matsuzaki H, Sugiyama T, Kawai M, Minabe Y, Takei N, Mori N. Related Articles, Links

Decreased Serum Levels of Epidermal Growth Factor in Adult Subjects with High-Functioning Autism.
Biol Psychiatry. 2006 Nov 20; [Epub ahead of print]
PMID: 17123472 [PubMed - as supplied by publisher]

□ 3: Grant CM, Apperly I, Oliver C. Related Articles, Links

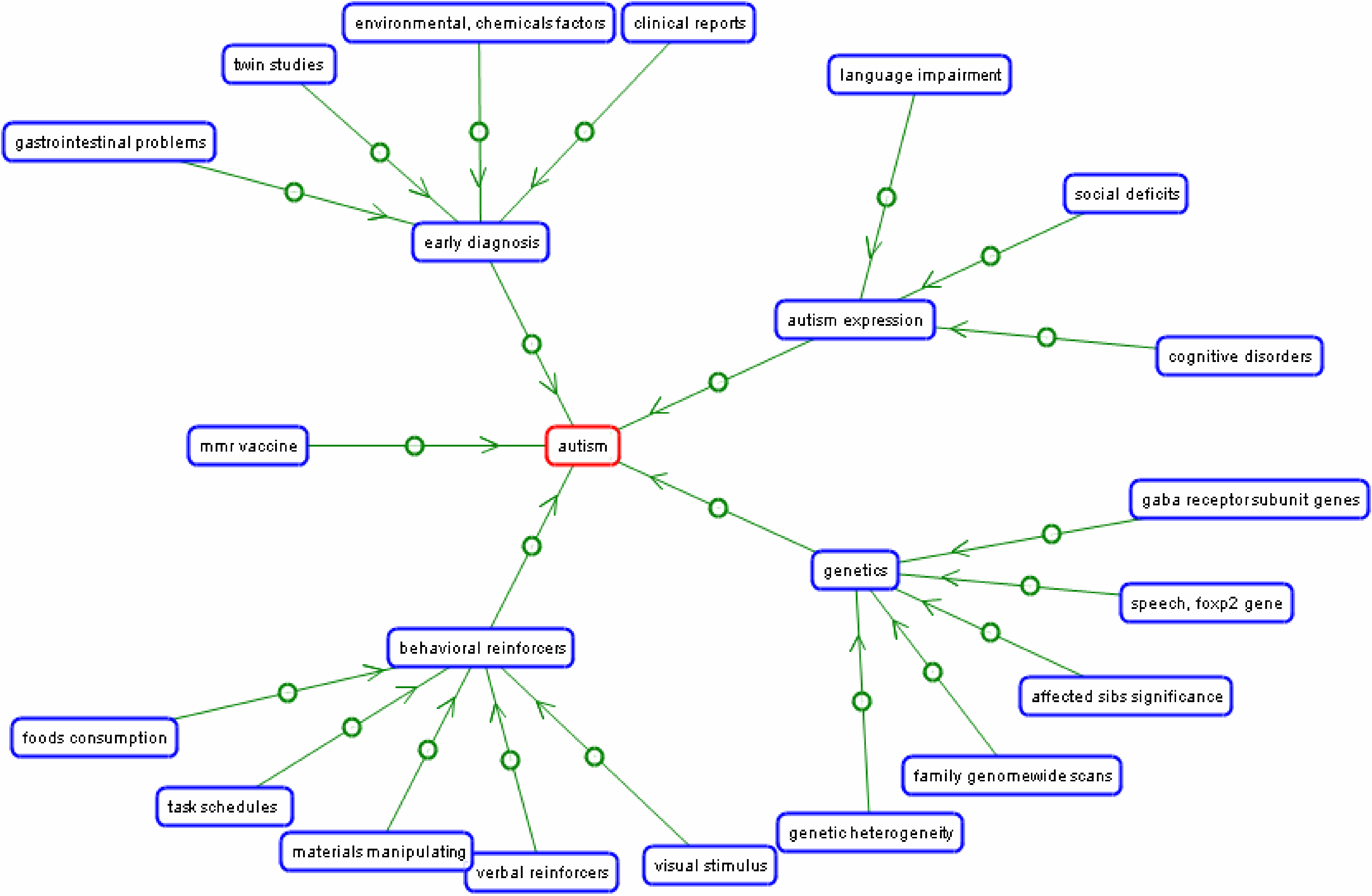
Is Theory of Mind Understanding Impaired in Males with Fragile X Syndrome?
J Abnorm Child Psychol. 2006 Nov 23; [Epub ahead of print]
PMID: 17123170 [PubMed - as supplied by publisher]

Data mining: Identification of domain structure I.

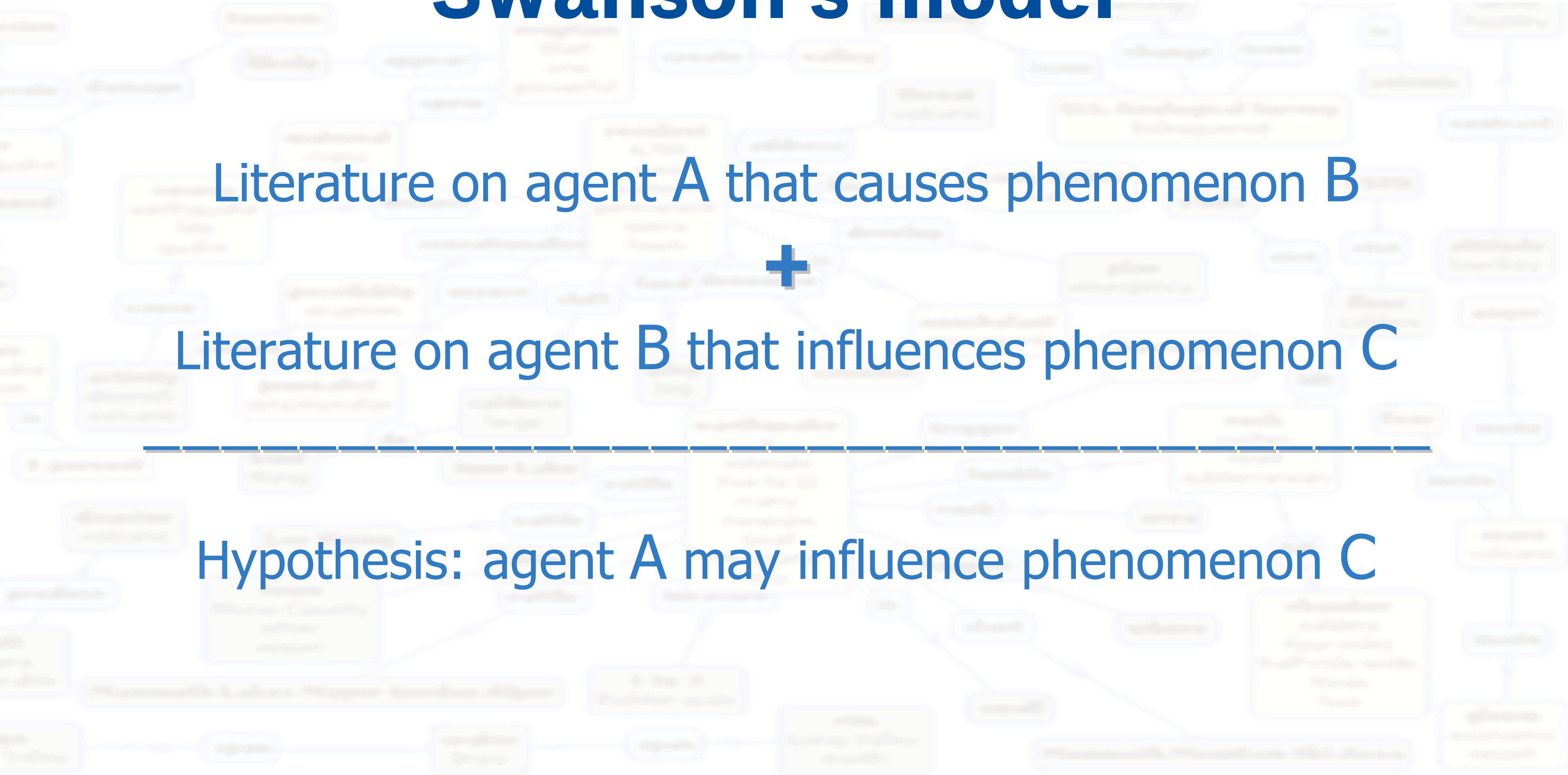


DEPARTMENT OF KNOWLEDGE TECHNOLOGIES
Jožef Stefan Institute

Data mining: Identification of domain structure II.



Hypothesis generation: Swanson’s model



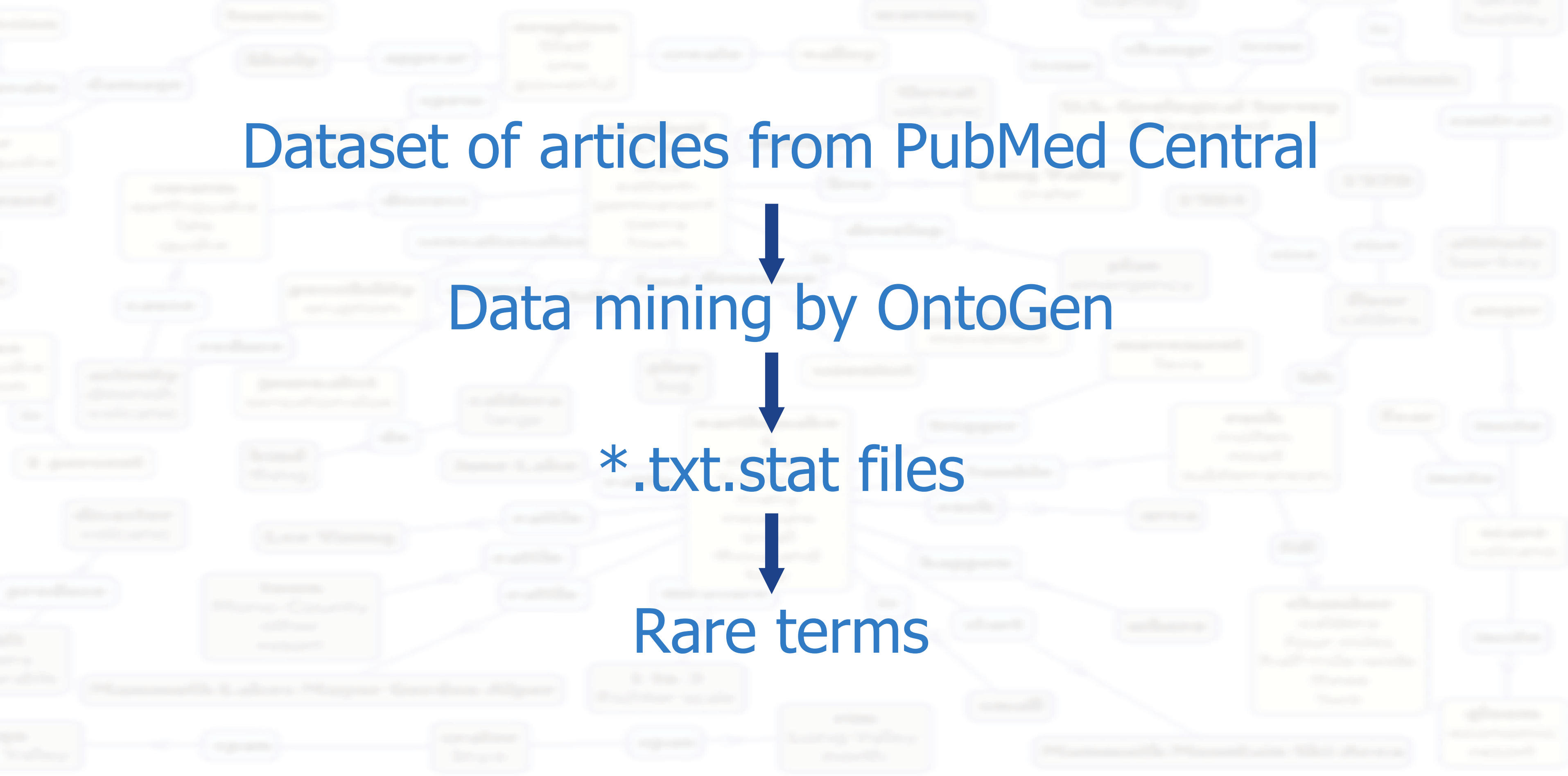
Literature on agent A that causes phenomenon B

+

Literature on agent B that influences phenomenon C

Hypothesis: agent A may influence phenomenon C

Investigation for rare data relations



Dataset of articles from PubMed Central

↓

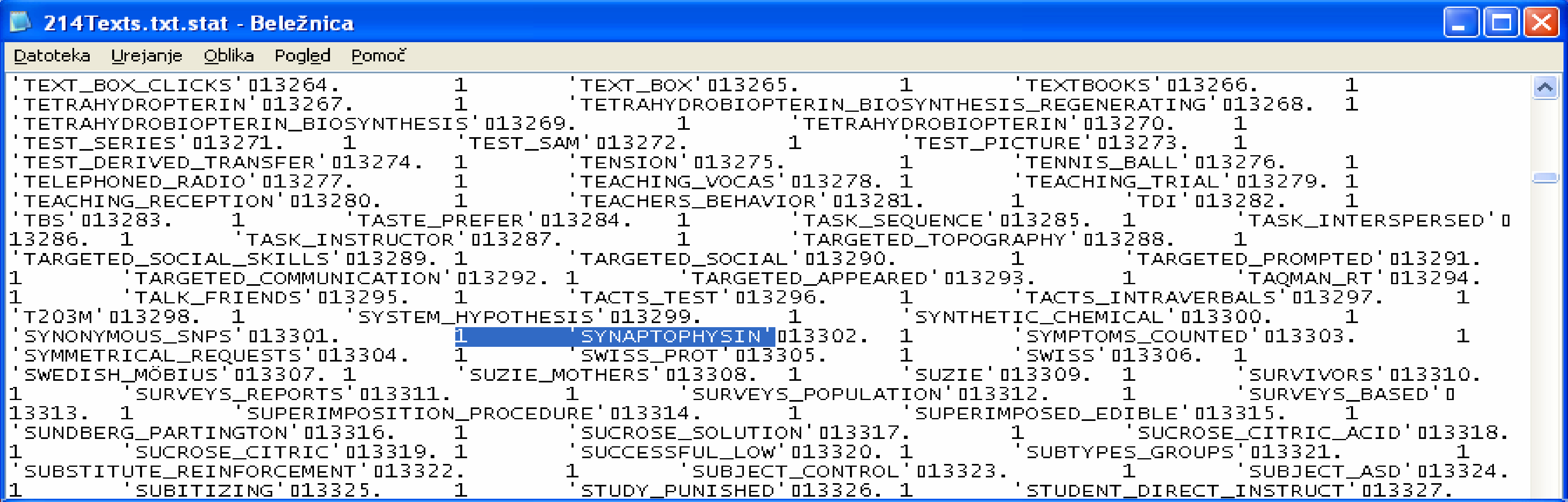
Data mining by OntoGen

↓

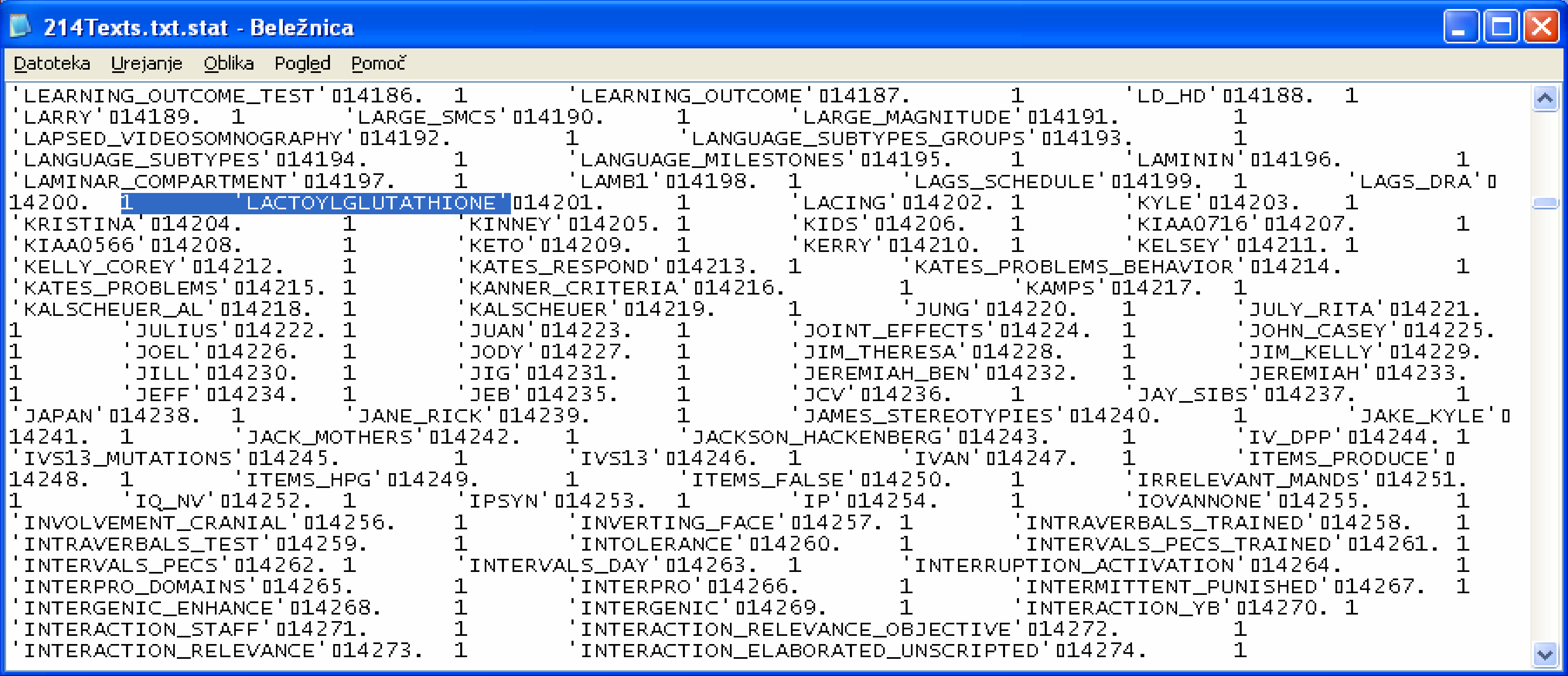
*.txt.stat files

↓

Rare terms

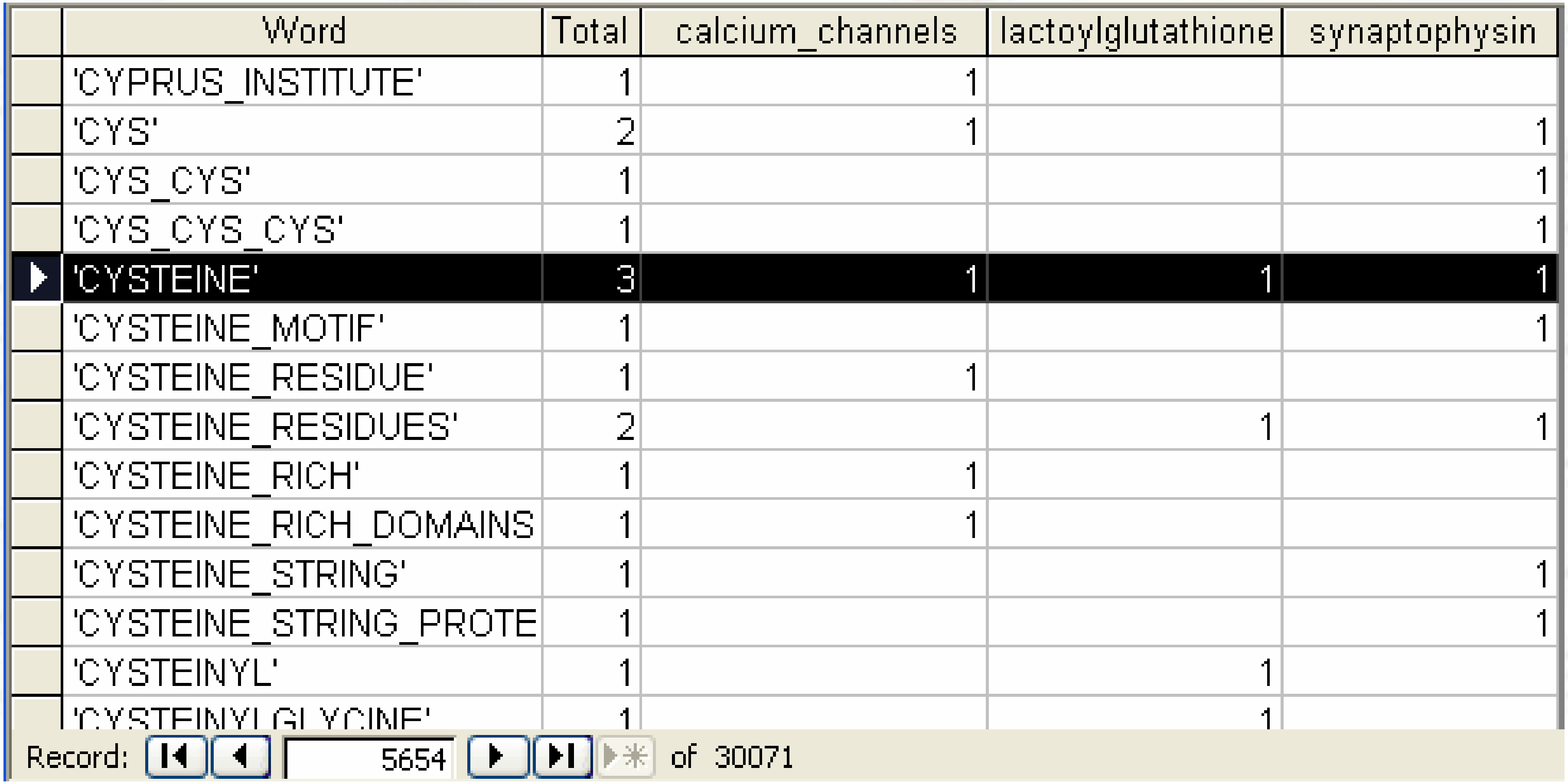


```
'TEXT_BOX_CLICKS'□13264.        1       'TEXT_BOX'□13265.       1       'TEXTBOOKS'□13266.      1
'TETRAHYDROPTERIN'□13267.       1       'TETRAHYDROBIOPTERIN_BIOSYNTHESIS_REGENERATING'□13268.  1
'TETRAHYDROBIOPTERIN_BIOSYNTHESIS'□13269.       1       'TETRAHYDROBIOPTERIN'□13270.    1
'TEST_SERIES'□13271.    1       'TEST_SAM'□13272.       1       'TEST_PICTURE'□13273.   1
'TEST_DERIVED_TRANSFER'□13274.  1       'TENSION'□13275.        1       'TENNIS_BALL'□13276.    1
'TELEPHONED_RADIO'□13277.       1       'TEACHING_VOCAS'□13278. 1       'TEACHING_TRIAL'□13279. 1
'TEACHING_RECEPTION'□13280.     1       'TEACHERS_BEHAVIOR'□13281.      1       'TDI'□13282.    1
'TBS'□13283.    1       'TASTE_PREFER'□13284.   1       'TASK_SEQUENCE'□13285.  1       'TASK_INTERSPERSED'□
13286.  1       'TASK_INSTRUCTOR'□13287.        1       'TARGETED_TOPOGRAPHY'□13288.    1
'TARGETED_SOCIAL_SKILLS'□13289. 1       'TARGETED_SOCIAL'□13290.        1       'TARGETED_PROMPTED'□13291.
1       'TARGETED_COMMUNICATION'□13292. 1       'TARGETED_APPEARED'□13293.      1       'TAQMAN_RT'□13294.
1       'TALK_FRIENDS'□13295.   1       'TACTS_TEST'□13296.     1       'TACTS_INTRAVERBALS'□13297.     1
'T203M'□13298.  1       'SYSTEM_HYPOTHESIS'□13299.      1       'SYNTHETIC_CHEMICAL'□13300.     1
'SYNONYMOUS_SNPS'□13301.        1       'SYNAPTOPHYSIN'□13302.  1       'SYMPTOMS_COUNTED'□13303.       1
'SYMMETRICAL_REQUESTS'□13304.   1       'SWISS_PROT'□13305.     1       'SWISS'□13306.  1
'SWEDISH_MÖBIUS'□13307. 1       'SUZIE_MOTHERS'□13308.  1       'SUZIE'□13309.  1       'SURVIVORS'□13310.
1       'SURVEYS_REPORTS'□13311.        1       'SURVEYS_POPULATION'□13312.     1       'SURVEYS_BASED'□
13313.  1       'SUPERIMPOSITION_PROCEDURE'□13314.      1       'SUPERIMPOSED_EDIBLE'□13315.    1
'SUNDBERG_PARTINGTON'□13316.    1       'SUCROSE_SOLUTION'□13317.       1       'SUCROSE_CITRIC_ACID'□13318.
1       'SUCROSE_CITRIC'□13319. 1       'SUCCESSFUL_LOW'□13320. 1       'SUBTYPES_GROUPS'□13321.        1
'SUBSTITUTE_REINFORCEMENT'□13322.       1       'SUBJECT_CONTROL'□13323.        1       'SUBJECT_ASD'□13324.
1       'SUBITIZING'□13325.     1       'STUDY_PUNISHED'□13326. 1       'STUDENT_DIRECT_INSTRUCT'□13327.
```



```
'LEARNING_OUTCOME_TEST'□14186.  1       'LEARNING_OUTCOME'□14187.       1       'LD_HD'□14188.  1
'LARRY'□14189.  1       'LARGE_SMCS'□14190.     1       'LARGE_MAGNITUDE'□14191.        1
'LAPSED_VIDEOSOMNOGRAPHY'□14192.        1       'LANGUAGE_SUBTYPES_GROUPS'□14193.       1
'LANGUAGE_SUBTYPES'□14194.      1       'LANGUAGE_MILESTONES'□14195.    1       'LAMININ'□14196.        1
'LAMINAR_COMPARTMENT'□14197.    1       'LAMB1'□14198.  1       'LAGS_SCHEDULE'□14199.  1       'LAGS_DRA'□
14200.  1       'LACTOYLGLUTATHIONE'□14201.     1       'LACING'□14202. 1       'KYLE'□14203.   1
'KRISTINA'□14204.       1       'KINNEY'□14205. 1       'KIDS'□14206.   1       'KIAA0716'□14207.       1
'KIAA0566'□14208.       1       'KETO'□14209.   1       'KERRY'□14210.  1       'KELSEY'□14211. 1
'KELLY_COREY'□14212.    1       'KATES_RESPOND'□14213.  1       'KATES_PROBLEMS_BEHAVIOR'□14214.        1
'KATES_PROBLEMS'□14215. 1       'KANNER_CRITERIA'□14216.        1       'KAMPS'□14217.  1
'KALSCHEUER_AL'□14218.  1       'KALSCHEUER'□14219.     1       'JUNG'□14220.   1       'JULY_RITA'□14221.
1       'JULIUS'□14222. 1       'JUAN'□14223.   1       'JOINT_EFFECTS'□14224.  1       'JOHN_CASEY'□14225.
1       'JOEL'□14226.   1       'JODY'□14227.   1       'JIM_THERESA'□14228.    1       'JIM_KELLY'□14229.
1       'JILL'□14230.   1       'JIG'□14231.    1       'JEREMIAH_BEN'□14232.   1       'JEREMIAH'□14233.
1       'JEFF'□14234.   1       'JEB'□14235.    1       'JCV'□14236.    1       'JAY_SIBS'□14237.       1
'JAPAN'□14238.  1       'JANE_RICK'□14239.      1       'JAMES_STEREOTYPIES'□14240.     1       'JAKE_KYLE'□
14241.  1       'JACK_MOTHERS'□14242.   1       'JACKSON_HACKENBERG'□14243.     1       'IV_DPP'□14244. 1
'IVS13_MUTATIONS'□14245.        1       'IVS13'□14246.  1       'IVAN'□14247.   1       'ITEMS_PRODUCE'□
14248.  1       'ITEMS_HPG'□14249.      1       'ITEMS_FALSE'□14250.    1       'IRRELEVANT_MANDS'□14251.
1       'IQ_NV'□14252.  1       'IPSYN'□14253.  1       'IP'□14254.     1       'IOVANNONE'□14255.      1
'INVOLVEMENT_CRANIAL'□14256.    1       'INVERTING_FACE'□14257. 1       'INTRAVERBALS_TRAINED'□14258.   1
'INTRAVERBALS_TEST'□14259.      1       'INTOLERANCE'□14260.    1       'INTERVALS_PECS_TRAINED'□14261. 1
'INTERVALS_PECS'□14262. 1       'INTERVALS_DAY'□14263.  1       'INTERRUPTION_ACTIVATION'□14264.        1
'INTERPRO_DOMAINS'□14265.       1       'INTERPRO'□14266.       1       'INTERMITTENT_PUNISHED'□14267.  1
'INTERGENIC_ENHANCE'□14268.     1       'INTERGENIC'□14269.     1       'INTERACTION_YB'□14270. 1
'INTERACTION_STAFF'□14271.      1       'INTERACTION_RELEVANCE_OBJECTIVE'□14272.        1
'INTERACTION_RELEVANCE'□14273.  1       'INTERACTION_ELABORATED_UNSCRIPTED'□14274.      1
```

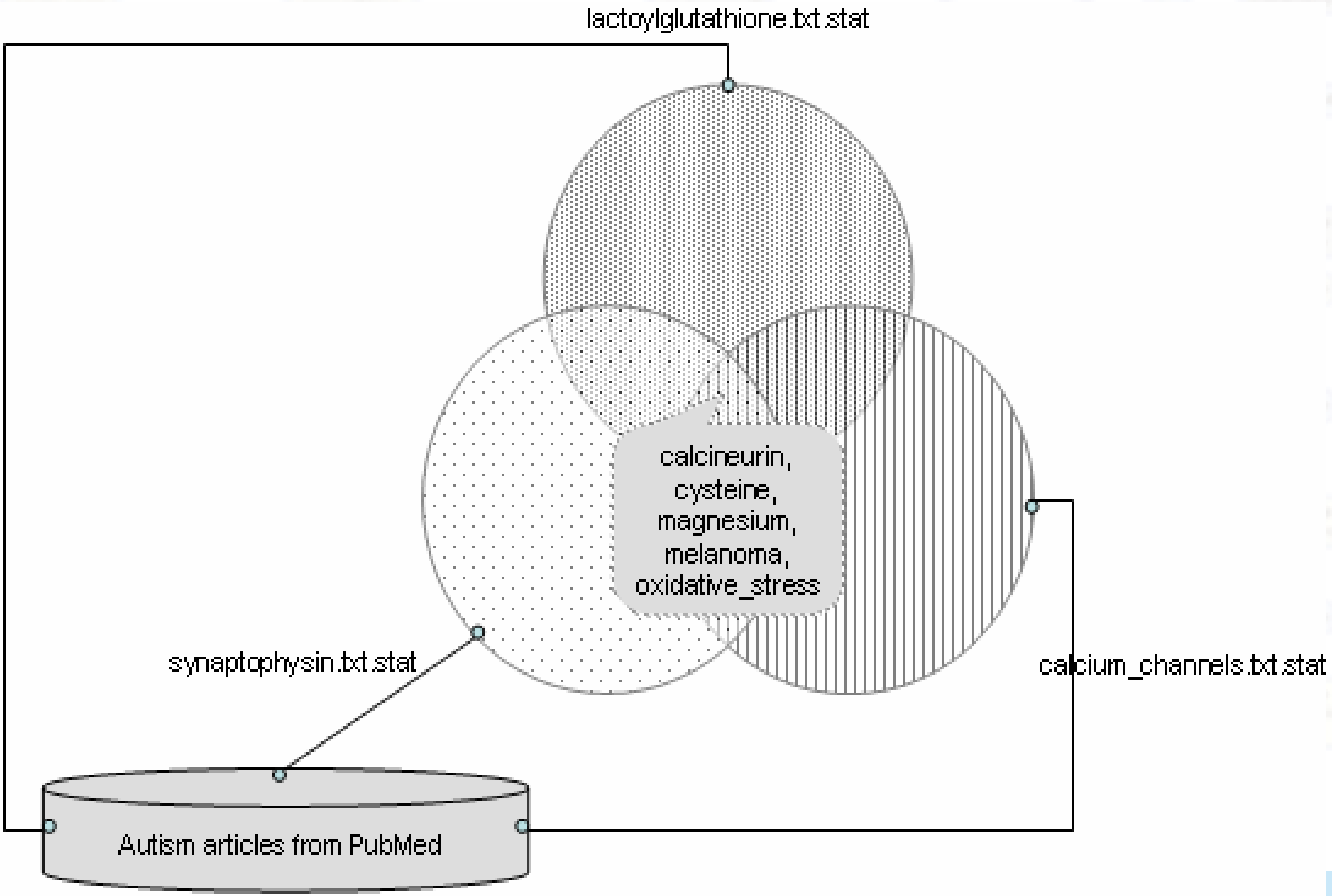
Connecting terms



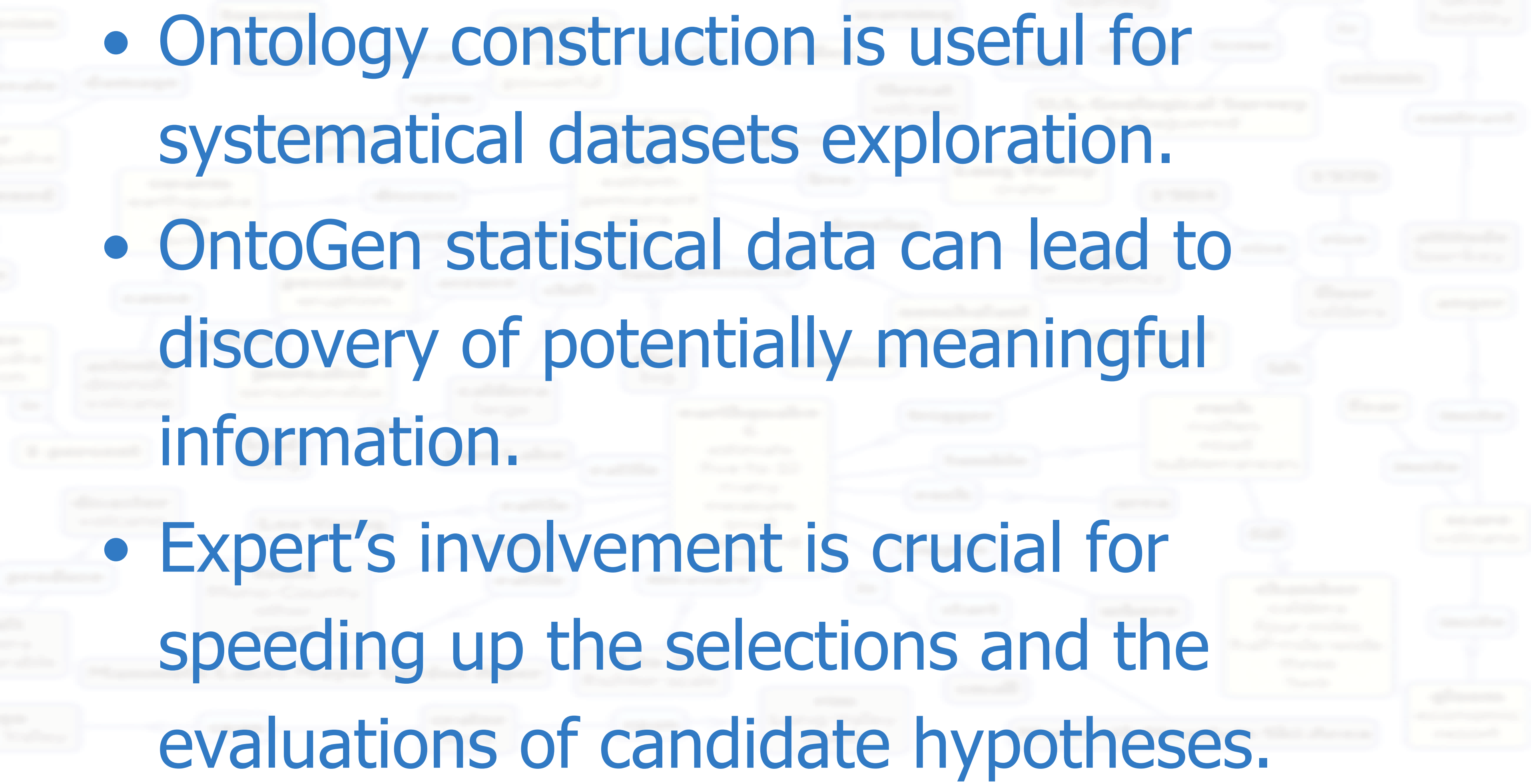
	Word	Total	calcium_channels	lactoylglutathione	synaptophysin
	'CYPRUS_INSTITUTE'	1	1		
	'CYS'	2	1		1
	'CYS_CYS'	1			1
	'CYS_CYS_CYS'	1			1
▶	'CYSTEINE'	3	1	1	1
	'CYSTEINE_MOTIF'	1			1
	'CYSTEINE_RESIDUE'	1	1		
	'CYSTEINE_RESIDUES'	2		1	1
	'CYSTEINE_RICH'	1	1		
	'CYSTEINE_RICH_DOMAINS	1	1		
	'CYSTEINE_STRING'	1			1
	'CYSTEINE_STRING_PROTE	1			1
	'CYSTEINYL'	1		1	
	'CYSTEINYLGLYCINE'	1		1	

Record: 5654 of 30071

Connecting sets of literature


lactoylglutathione.txt.stat
calcineurin,
cysteine,
magnesium,
melanoma,
oxidative_stress
synaptophysin.txt.stat
calcium_channels.txt.stat
Autism articles from PubMed

Conclusions



- Ontology construction is useful for systematical datasets exploration.
- OntoGen statistical data can lead to discovery of potentially meaningful information.
- Expert’s involvement is crucial for speeding up the selections and the evaluations of candidate hypotheses.