Exploring ontology metrics in the biomedical domain

N. Manouselis, M.A. Sicilia, **Daniel Rodríguez**
University of Alcalá

Outline

- Background
- Ontologies
  - http://www.obofoundry.org/
- Ontology metrics
- Empirical results
- Conclusions and future work

Introduction

- Initial exploratory study on measuring a set of ontologies published in the Open Biomedical Ontologies (OBO) repository
  - Collected using a common topical criteria and maintained with a coherent set of tools.
- The study has been carried out by implementing and using an open source software framework for computing ontology metrics expressed in the Ontology Web Language (OWL).
- The overall statistics are reported, along with an exploratory study on potential categories of ontologies with diverging characteristics for which different metric interpretation or different quality criteria could be appropriate.
  - http://www.obofoundry.org/

Ontometrics Framework

- OWL ontologies are composed of
  - (i) *classes* that can be nested as sets of individuals
  - (ii) *individuals* as instances of classes, i.e., objects of the domain and
  - (iii) *properties* as binary relations between individuals. It is also possible to specify property domains, cardinality ranges and reasoning on ontologies.
- From these basic elements a number of authors have proposed metrics to measure the quality of ontologies.
- We implemented a Java framework based on the Protégé API
  - http://www.cc.uah.es/ie/software/OntoMetrics.zip
- Currently upgrading to OWL2 API
  - http://owlapi.sourceforge.net/

Metrics Implemented

- No. of Classes (*noc*)
  - count of the number of classes contained in the ontology.
- No. of Instances (*noi*)
  - count of the number of instances contained in the ontology.
- No. of Properties (*nop*)
  - count of the number of properties contained in the ontology.
- *Number of Root Classes* metric (*norc*)
  - corresponds to the number of root classes (those without superclasses) explicitly defined.
- Number of Leaf Classes metric (nolc)
  - the sum of all leaf classes, i.e., those without subclasses, in an ontology

Metrics Implemented

- Average Population metric (ap)
  - measures the average distribution of instances across all classes.
- Class Richness metric (cr)
  - ratio between the number of classes that have instances divided by the total number of classes.
- Explicit Depth of Submission Hierarchy (dosh)
- Relationship Richness metric (rr)
  - ratio of the number of relationships defined in the schema divided by the sum of the number of subclasses.
- Inheritance Richness metric (ir)
  - the average number of subclasses per class

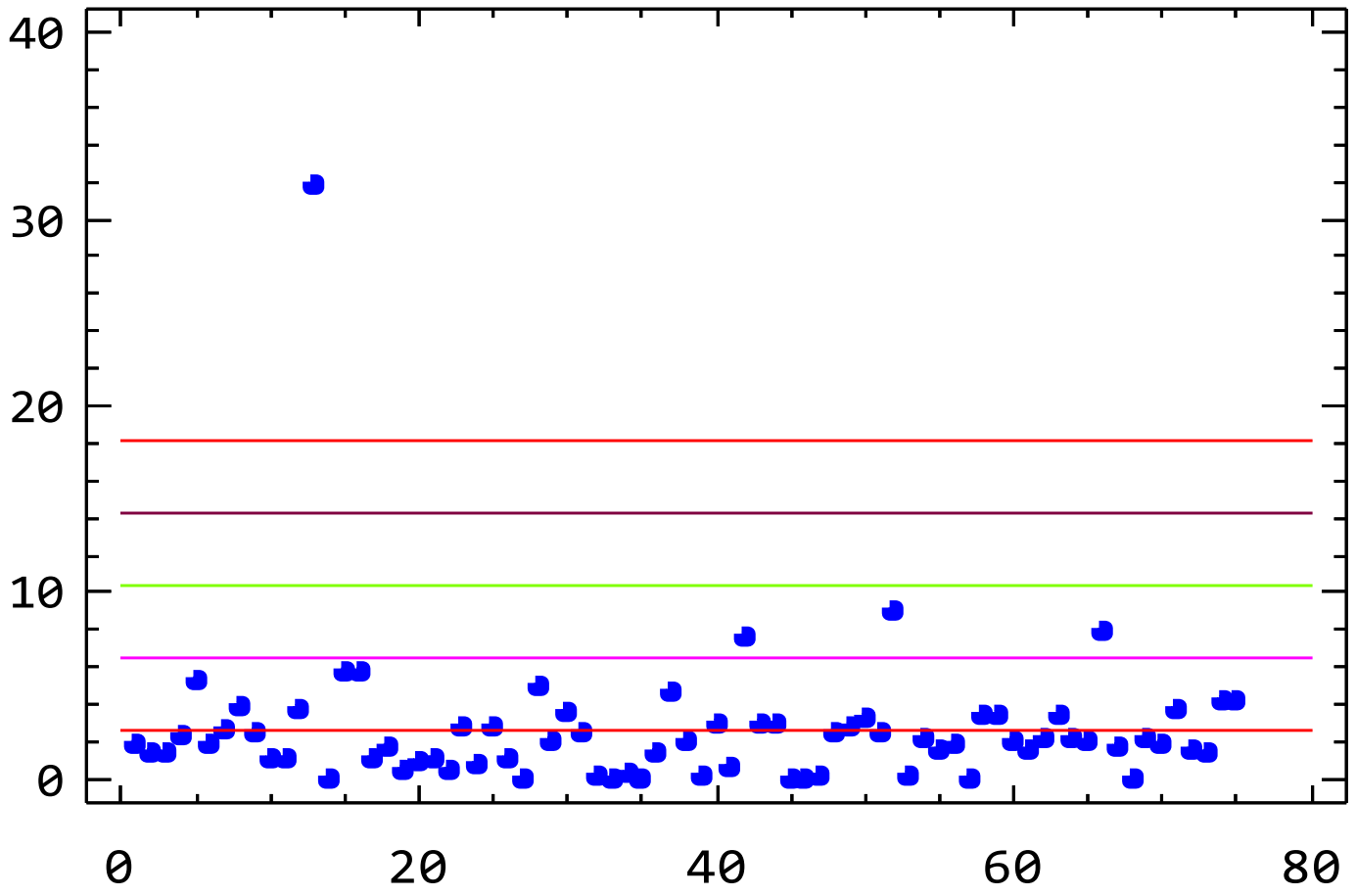
Descriptive Statistics

	ap	cr	dosh	lr	noc	noi	nolc	nop	norc	rr
Count	75	75	75	75	75	75	75	75	75	75
Avg	2.64	0.01	10.12	0.93	3169.75	11318.8	2490.52	15.41	496.23	0.44
Variance	15.16	0.00	37.27	0.09	8.37E+07	2.87E+09	5.36E+07	947.27	3.65E+06	0.12
StdDev	3.89	0.05	6.10	0.30	9148.88	53541.6	7318.05	30.78	1910.5	0.35
Min	0	0	1	0	34	0	21	0	1	0
Max	31.76	0.41	41	1.62	75529	455734	60858	194	13737	1
Range	31.76	0.41	40	1.62	75495	455734	60837	194	13736	1
Std Sk	20.82	25.98	10.00	-4.93	24.66	28.26	25.18	13.08	20.12	0.01
Std Kr	76.05	104.35	20.86	5.41	96.18	117.62	99.54	29.22	61.96	-2.43

Descriptive Statistics

- From the basic descriptive statistics for the set of 75 ontologies studied.
- Biomedical ontologies are relatively large ones when considering their t-boxes, with an average of more than 3000 explicitly declared ones.
  - However, there are some that are much smaller.
- From the distribution of classes, instances and properties in the ontologies contrasted with a normal curve.
- From the histograms, the three basic measures are distributed so that there are many ontologies with few elements.

Outlier Detection



Outlier Detection

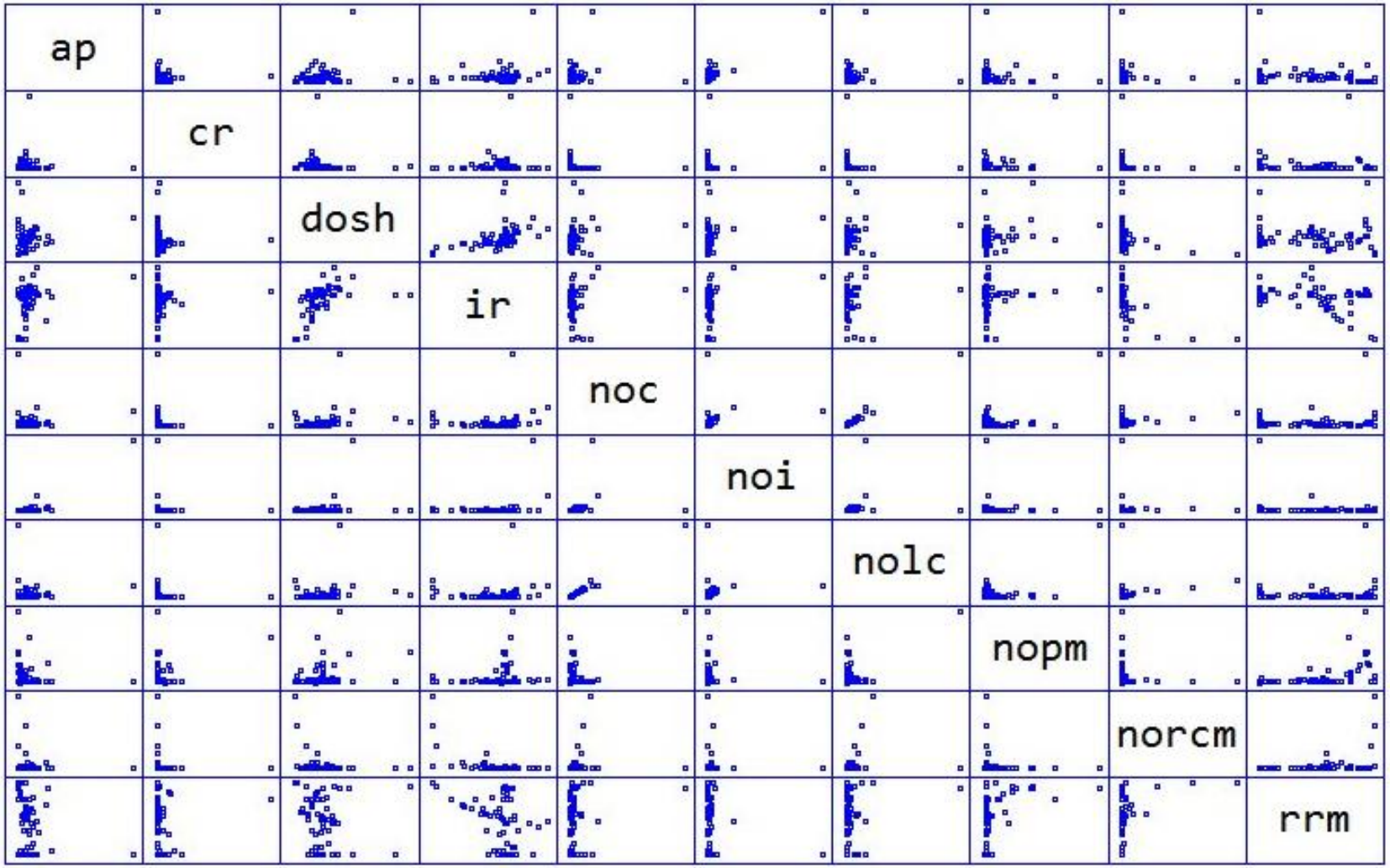
- These were the NCI Thesaurus (ncithesaurus.owl) and the disease_ontology.owl.
  - Both are extremely large ontologies (140MB and 150Mbytes respectively).
  - For example, using the *ap* (average population) variable, the most extreme value corresponds to the NCI thesaurus which is 7.48 standard deviations from the mean.
- A closer look into these ontologies reveals that these are special ontologies.
  - The NCI thesaurus is used to define a vocabulary of the cancer domain and related diseases, and not a formal ontology in an strict sense. The second one also defines a vocabulary of human diseases based on previous thesauri and terminologies.
- These can therefore be considered special ontologies that deserve separate examination.

Correlation Analysis

	ap	cr	dosh	ir	noc	noi	nolc	nop	norc	rr
		-0.027	0.141	0.24•	0.084	0.9••	0.051	-0.138	-0.089	-0.208
ap	—	0.82	0.227	0.038	0.473	0	0.663	0.238	0.446	0.074
			-0.094•	0.064	-0.094•	-0.059•	-0.093	0.356•	-0.074	0.036
cr		—	0.422	0.584	0.422	0.613	0.428	0.002	0.526	0.761
				0.448•	0.224	0.24•	0.185	0.313•	-0.263•	-0.02
dosh			—	0.000	0.054	0.038	0.112	0.006	0.023	0.868
					0.087	0.248•	0.024	0.14	-0.619•	-0.438•
ir				—	0.458	0.032	0.839	0.230	0	0
						0.192	0.995••	0.6•	0.153	0.174
noc					—	0.099	0	0	0.191	0.135
							0.151	-0.069	-0.024	-0.149
noi						—	0.196	0.554	0.840	0.202
								0.603•	0.211	0.2
nolc							—	0	0.07	0.085
									-0.1	0.474
nop								—	0.393	0
										0.338•
norc									—	0.003

Correlation Analysis

- Those *p-values* below 0.05 are represented with a "•" and indicate statistically significant non-zero correlations at the 95% confidence level.
- For the OBO repository there is a couple of pairs that are very positively correlated (marked with"••"):
  - (i) *noi* (no. of instances) and *ap* (average population),
  - (ii) *noc* (number of classes) and *nolc* (number of leaf classes).
- This suggests that ontologies are in general quite flat and most classes contain single instances.


ap
cr
dosh
ir
noc
noi
nolc
nopm
norcm
rrm

Factor Analysis

- The relationship between the rest of the metrics were contrasted using factor analysis with 3 components.
- The first component is characterized by high values in *dosh* and *ir*, that roughly measure depth and breadth of the ontology hierarchy.
  - Not surprisingly, this component is negatively correlated with *norc*, i.e. these ontologies tend to have fewer roots of hierarchy trees.
- The second component is characterized by a high correlation with the number of classes and properties and also with relationship richness that relates both of them.
- The third component is correlated with the number of instances.

Factor Analysis

	Component		
	1	2	3
cr	.068	.217	-.753
dosh	.666	.285	.230
ir	.893	-.122	-.088
noc	.210	.728	.371
noi	.394	-.007	.561
nop	.259	.891	-.272
norc	-.722	.238	.371
rr	-.433	.704	-.036

Conclusions and Future Work

- Implemented a framework for ontology metrics
- Preliminary study in a large number of ontologies

- Future work.
  - Reimplement those metrics with the new OWL API.
  - Further metrics with further ontologies and the ontoration
    - http://swoogle.umbc.edu/

- Questions?