Methods for Intrinsic Evaluation of Links in the Web of Data

Cristina Sarasua, Steffen Staab, Matthias Thimm

ESWC 2017

WeST
People and Knowledge Networks

Institute for Web Science and Technologies · University of Koblenz-Landau, Germany

"The Semantic Web isn't just about putting data on the Web. It is about making links, so that a person or machine can explore the Web of Data."

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include **links** to other URIs, so that they can **discover more things.**

[Berners-Lee, 2006]

Current Link Analyses

In- and Out-degree: “Bibsonomy has links to external 91 data sets.”

Overall link predicate usage: “owl:sameAs as the most widely used linking predicate.” [Schmachtenberg et al., 2014]

Link count: “The Eurostat data set has owl:sameAs 149 links to DBpedia.” [CKAN][Ermilov et al., 2013][Hogan et al.,2012]

Descriptive Statistics

Symmetry: “ the symmetry of entity links varies between different pairs of datasets.” [Hu et al., 2015 on Life sciences data sets]

Transitivity: “ the transitivity of entity links is often topic-dependent.” [Hu et al., 2015 on Life sciences data sets]

Link Properties

Current Link Quality Assessment Methods

Deadlinks: “we found 302,855,189 unverified links, and 12,430,800 dead links.”
[Neto et al., 2016]

Availability

Network measures and link properties count as proxy: “centrality, clustering coefficient, , sameAs chains are shown to be partially effective at detecting semantically correct / incorrect links.”
[Guéret et al., 2012]

Crowdsourcing: “hybrid methods can improve semantic accuracy.”
[Demartini et al., 2012, Sarasua et al.,2012, Acosta et al., 2013]

Semantic Accuracy

Entity Connectivity: “50 % of the entities in the source data set contain links to external entities.”
[Albertoni et al., 2013]

Completeness

Current Link Quality Assessment Methods

Deadlinks: "we found 302,855,189 unverified links, and 12,430,800 dead links."

[Neto et al., 2016]

Availability

Network measures and link properties count as proxy: "centrality, clustering coefficient, sameAs chains are shown to be partially effective at detecting semantically correct/incorrect links."

Crowdsourcing: "hybrid methods can improve semantic accuracy."
[Demartini et al., 2012]
[Sarasua et al.,2012]
[Acosta et al., 2013]

Semantic Accuracy

Interlinking/Connectivity: "50 % of the entities in the source data set contain links to external entities."

[Albertoni et al., 2013]

Completeness

To what extent do existing links add "more things" to the source entities?

Why?

Data Publisher
(predisposed to improve her links)

Help

- to understand the impact of existing links
- to spot weak points
- with guidance for improving existing links

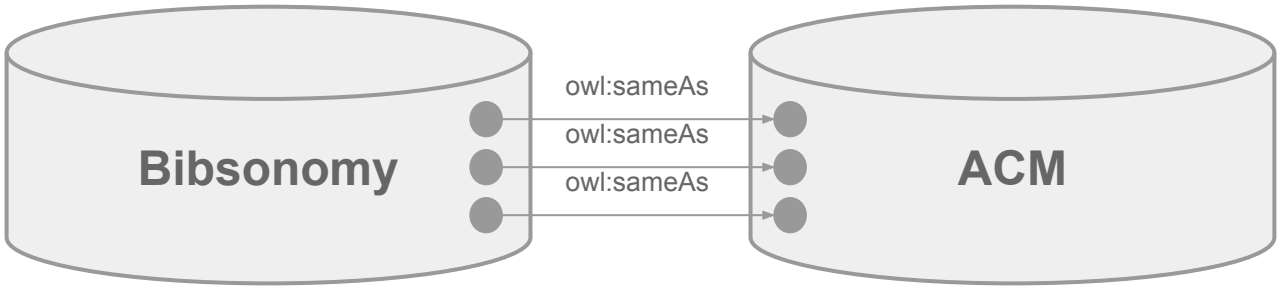
Encourage iterative maintenance

Our Task

Given a data set D containing the interlinking I

- ➔ compare D and D\ I and
- ➔ analyse the value that I gives to the source data

In terms of the principles for data interlinking in the Web of Data.

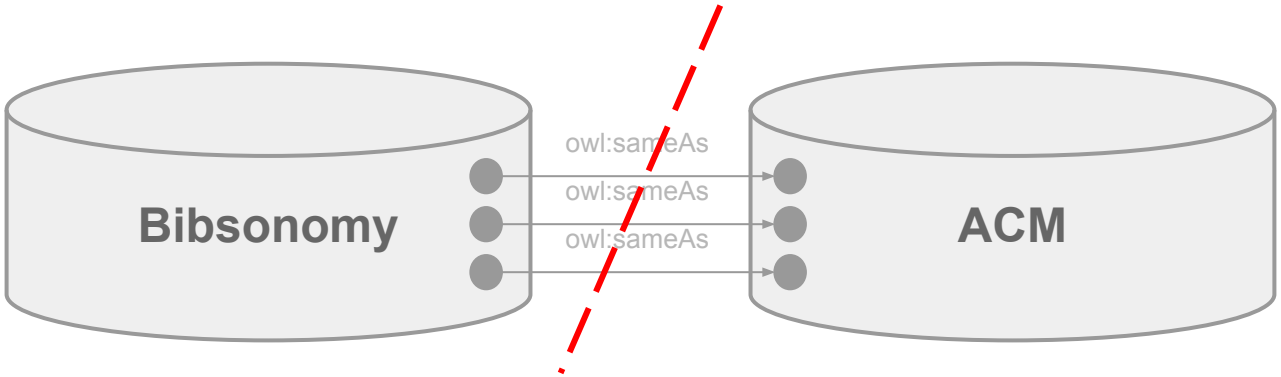


Our Task

Given a data set D containing the interlinking I

- ➔ compare D and D\ I and
- ➔ analyse the value that I gives to the source data

In terms of the principles for data interlinking in the Web of Data.

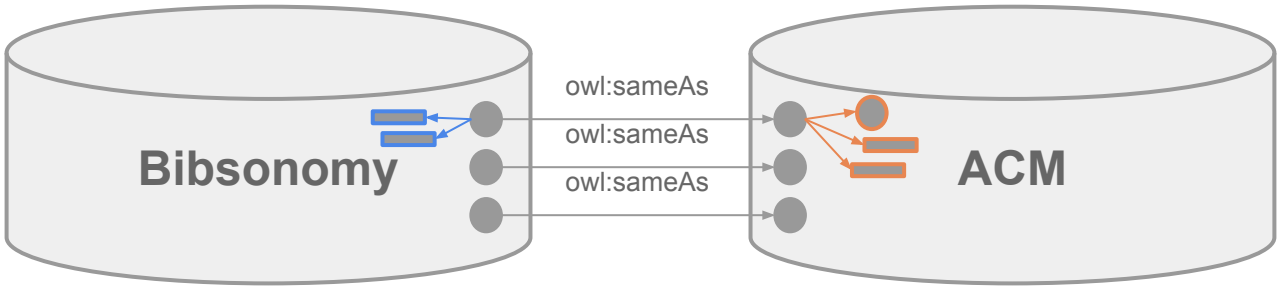


Our Task

Given a data set D containing the interlinking I

- ➔ compare D and D\ I and
- ➔ analyse the value that I gives to the source data

In terms of the principles for data interlinking in the Web of Data.

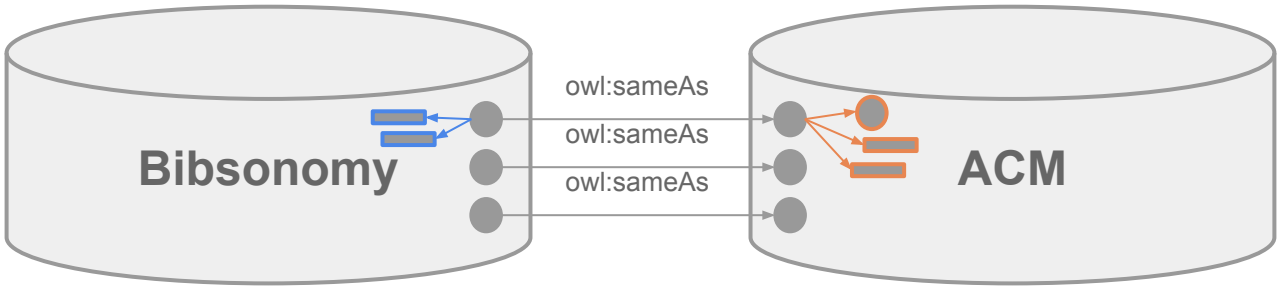


Our Task

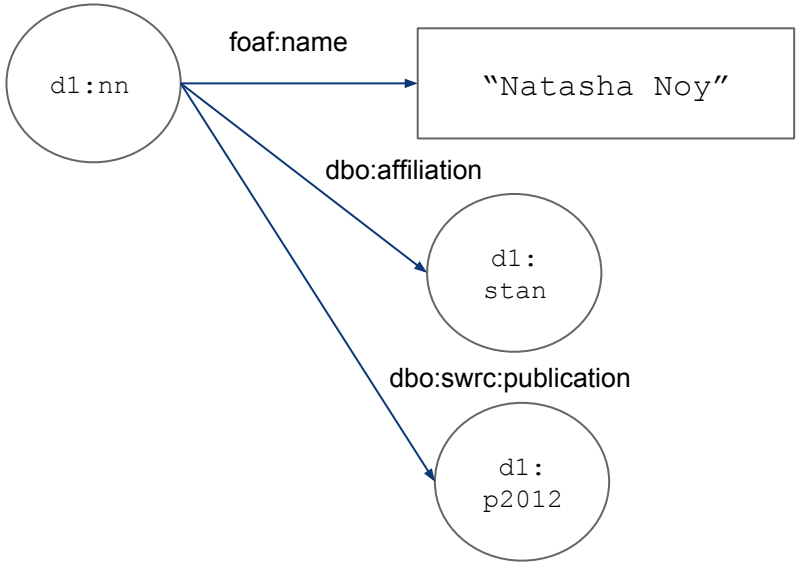
Given a data set D containing the interlinking I

- ➔ compare D and D\ I and
- ➔ analyse the value that I gives to the source data

In terms of the principles for data interlinking in the Web of Data.



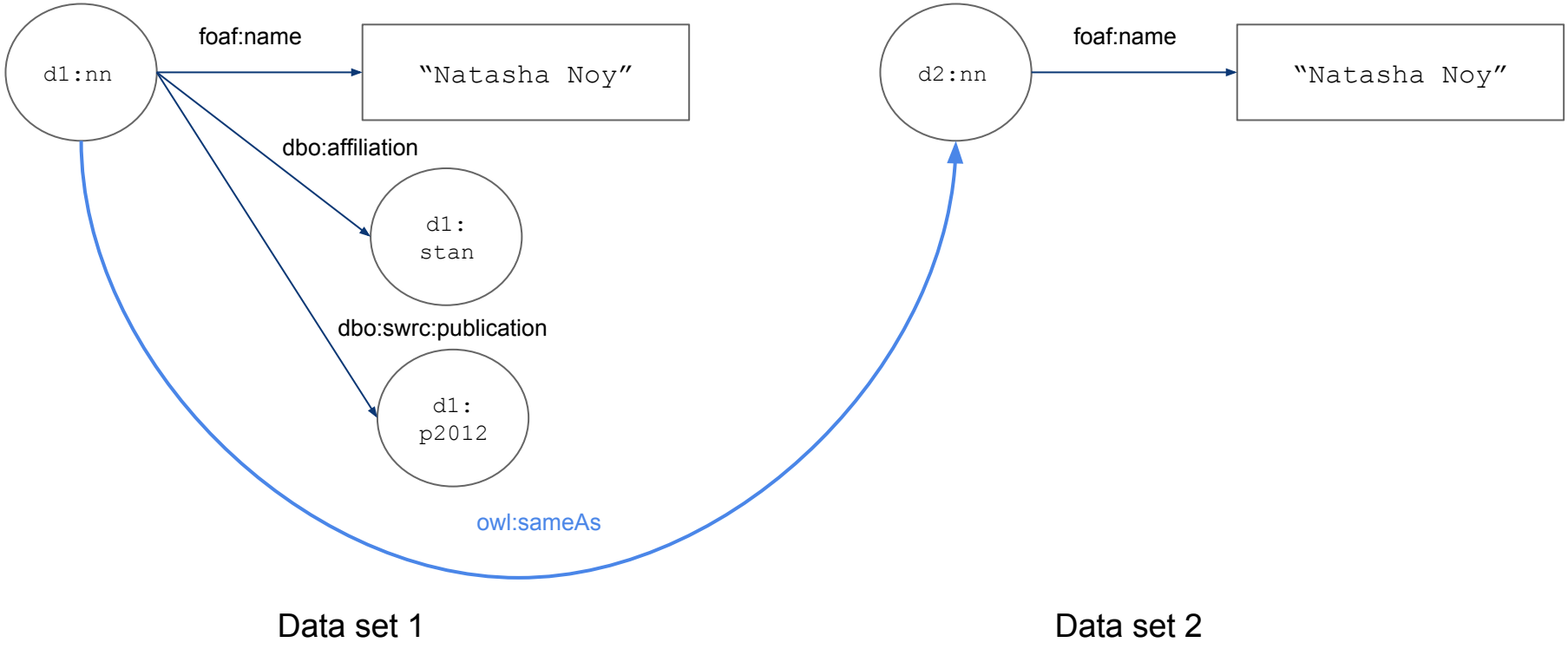
Principles



Data set 1

Principles

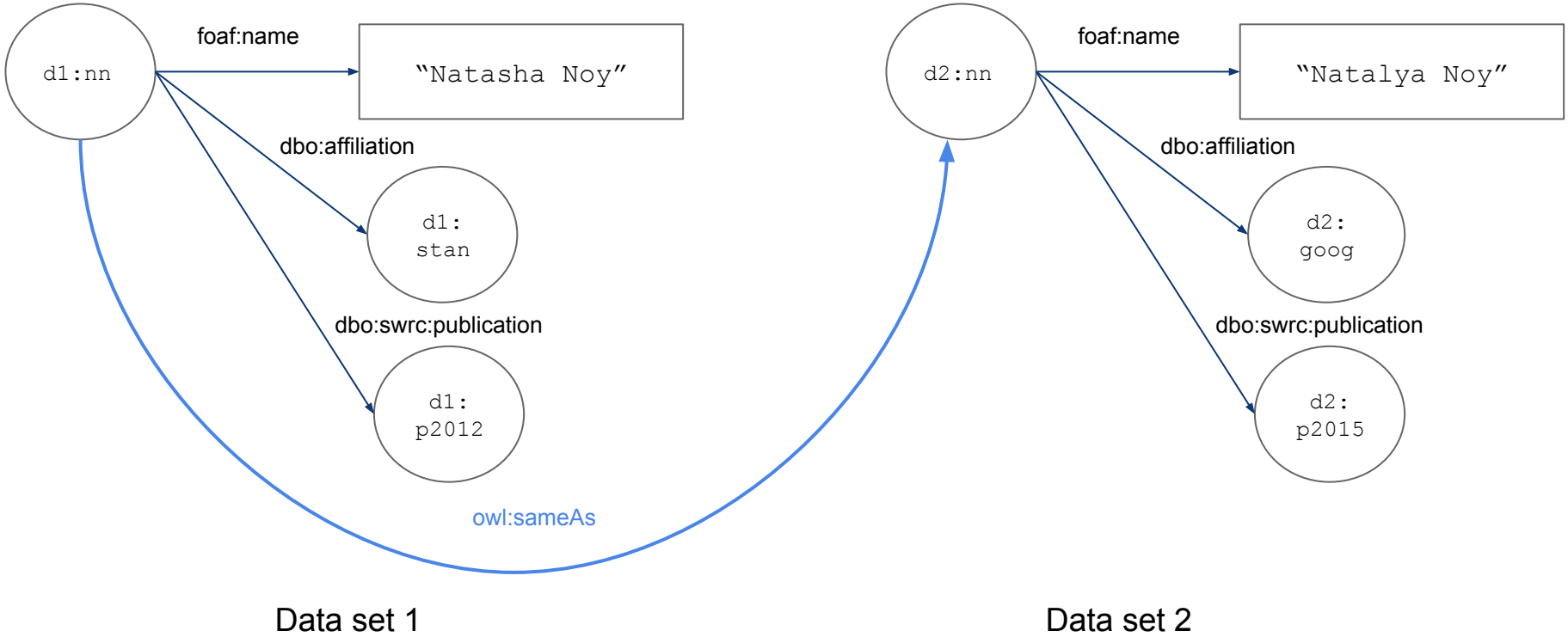
Extend entity description (P1)



Principles

Extend entity description (P1)

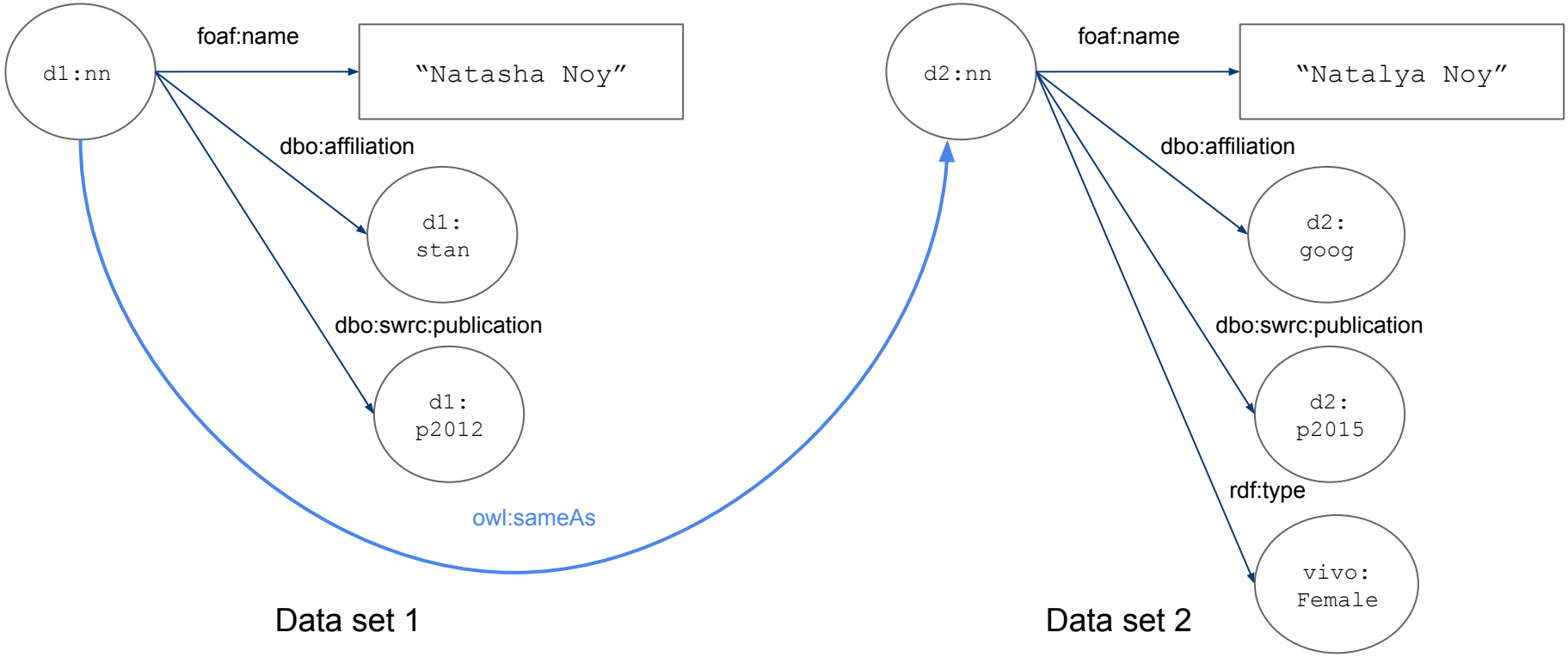
Better!



Principles

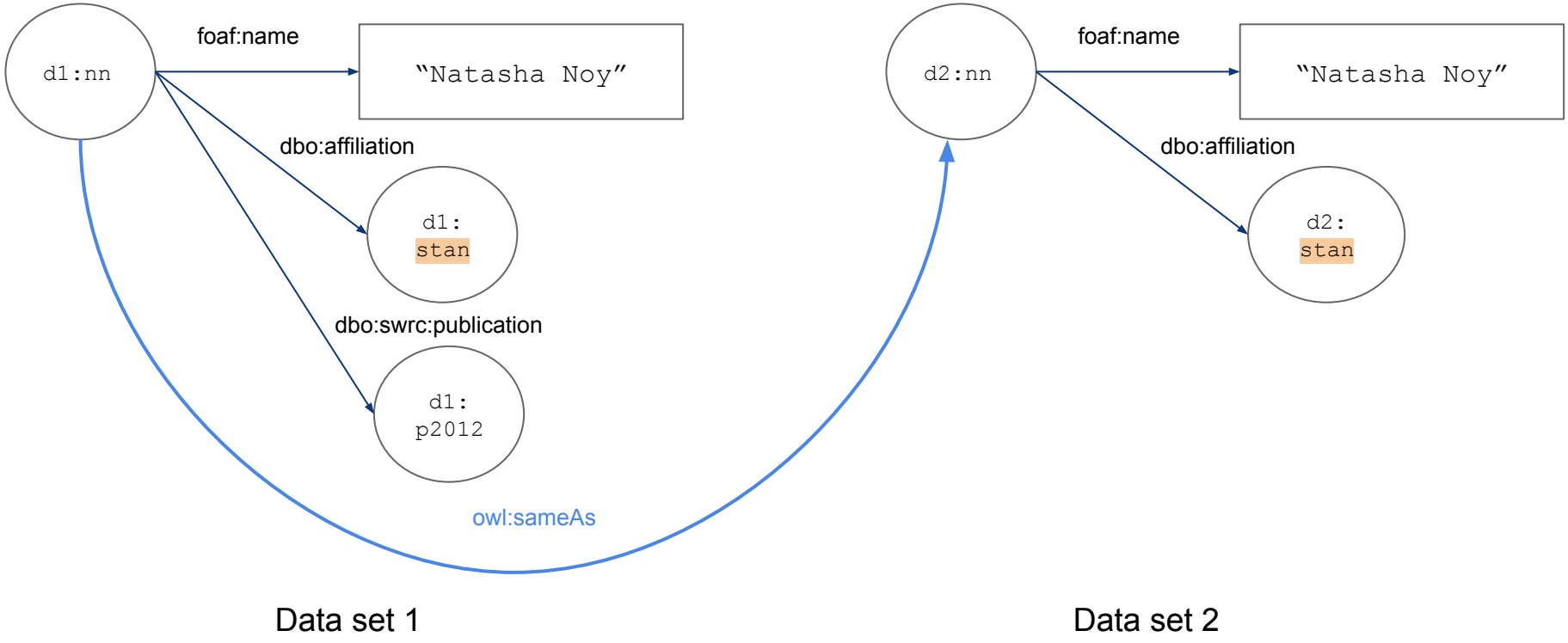
Extend entity description (P1)

Even Better!



Principles

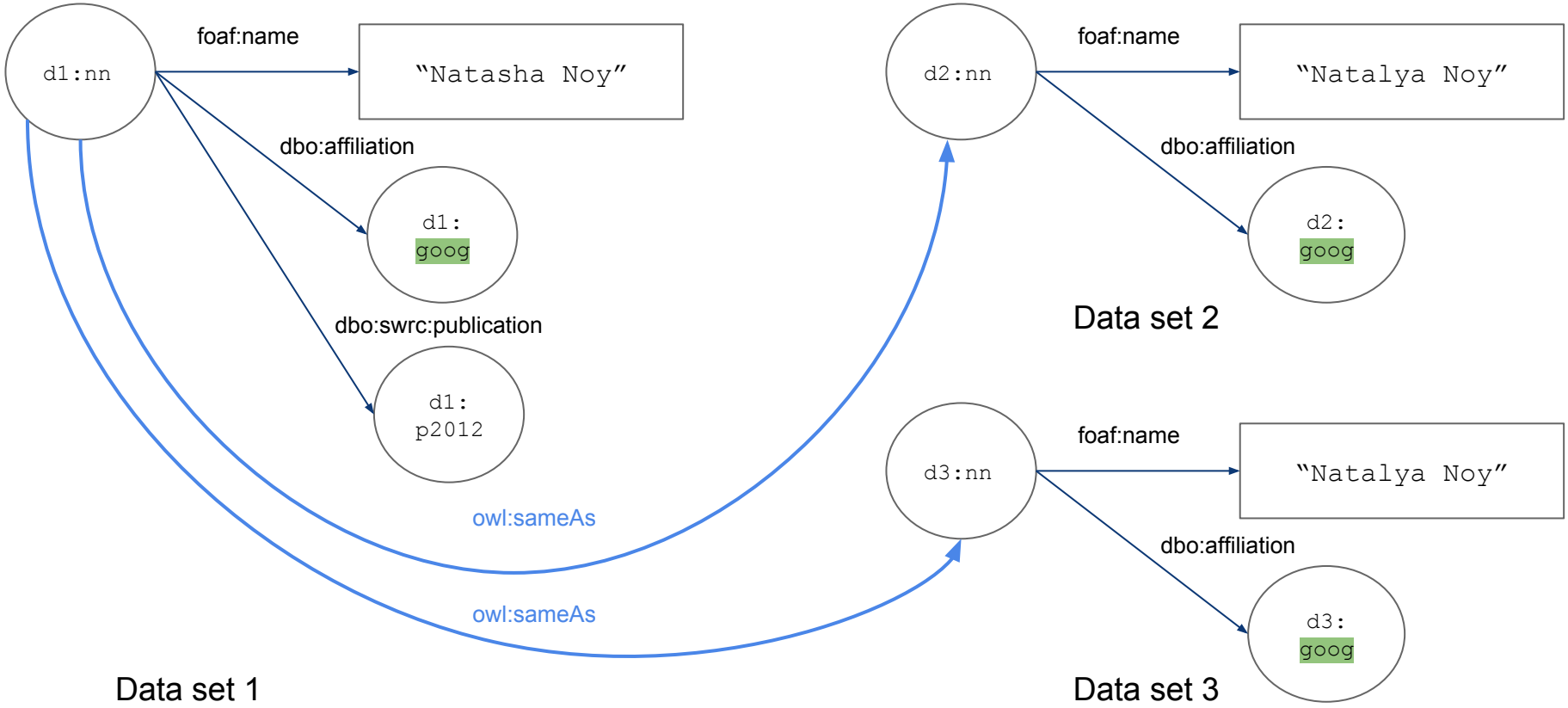
Extend entity connectivity (P2)



Principles

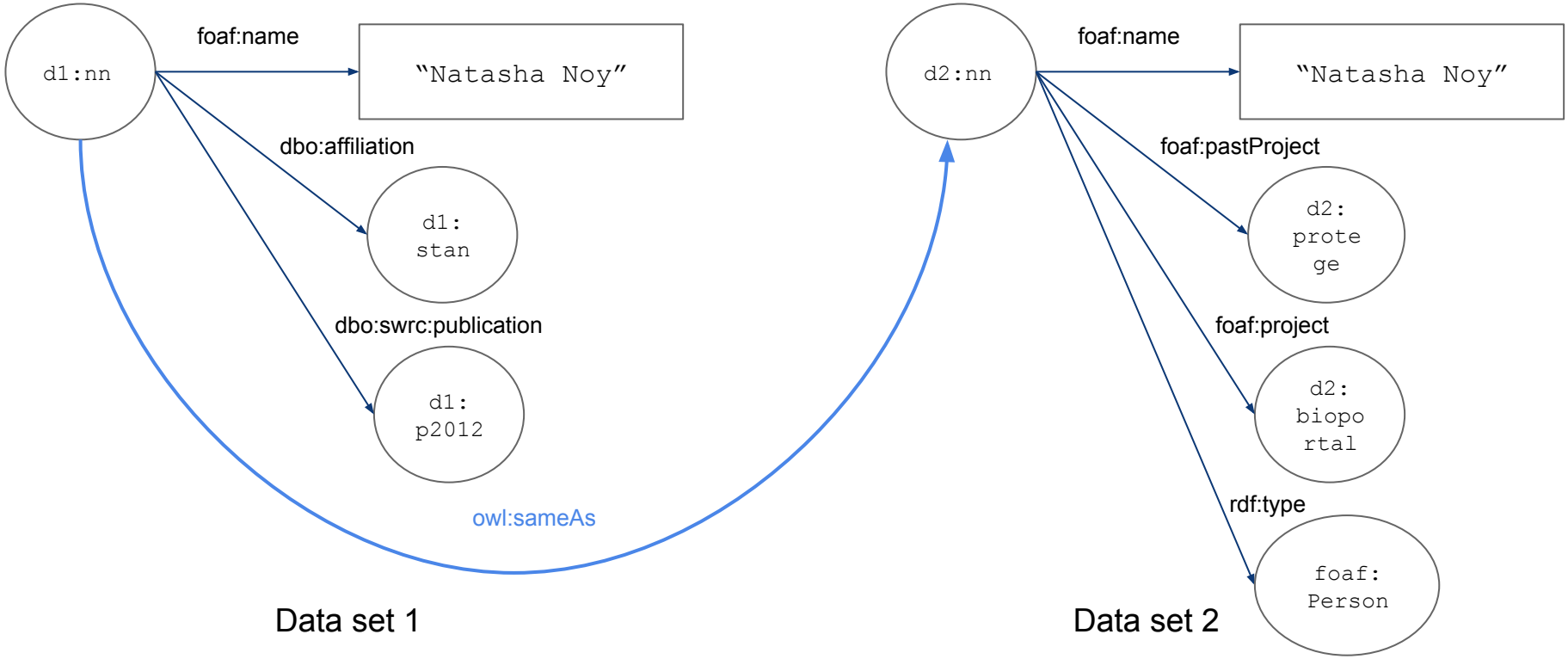
Extend entity connectivity (P2)

Better!



Principles

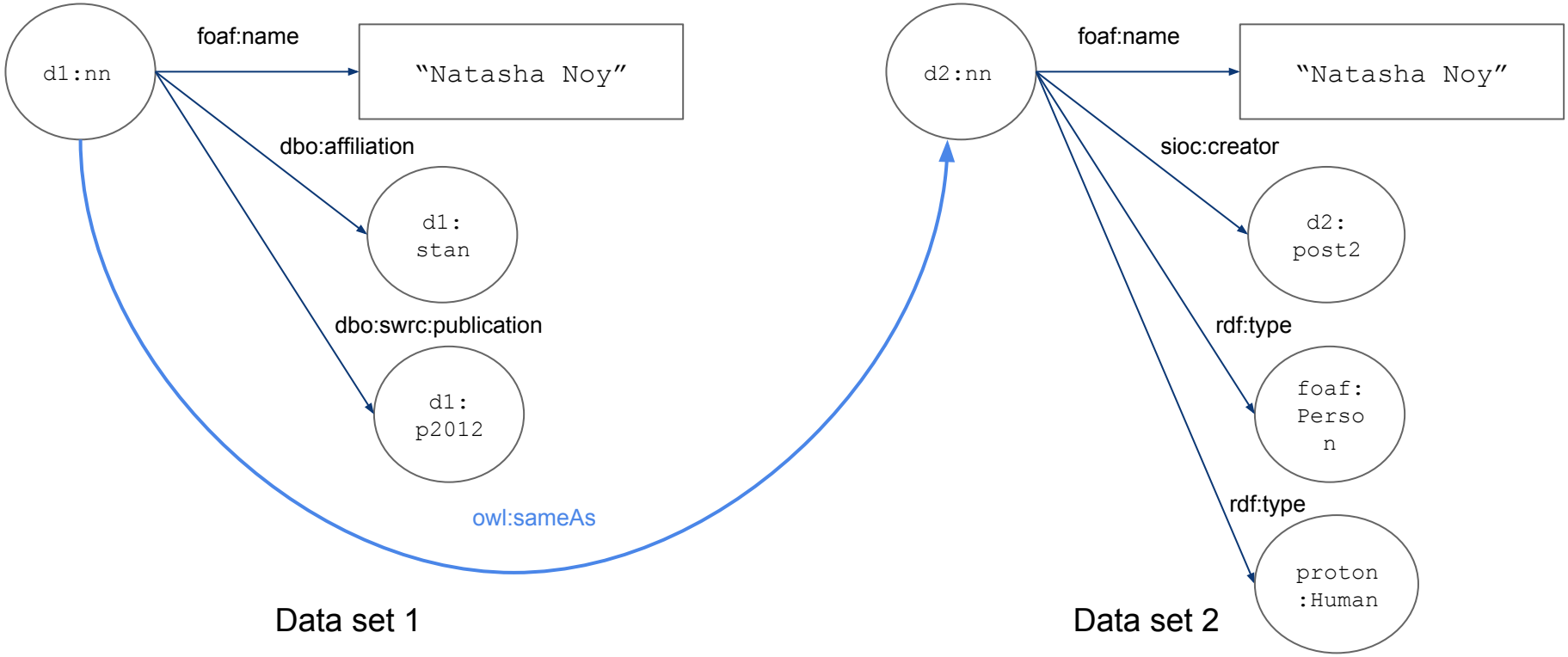
Increase number of vocabularies used (P3)

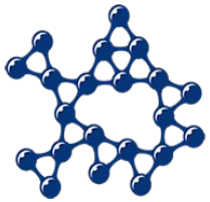


Principles

Increase number of vocabularies used (P3)

Better!



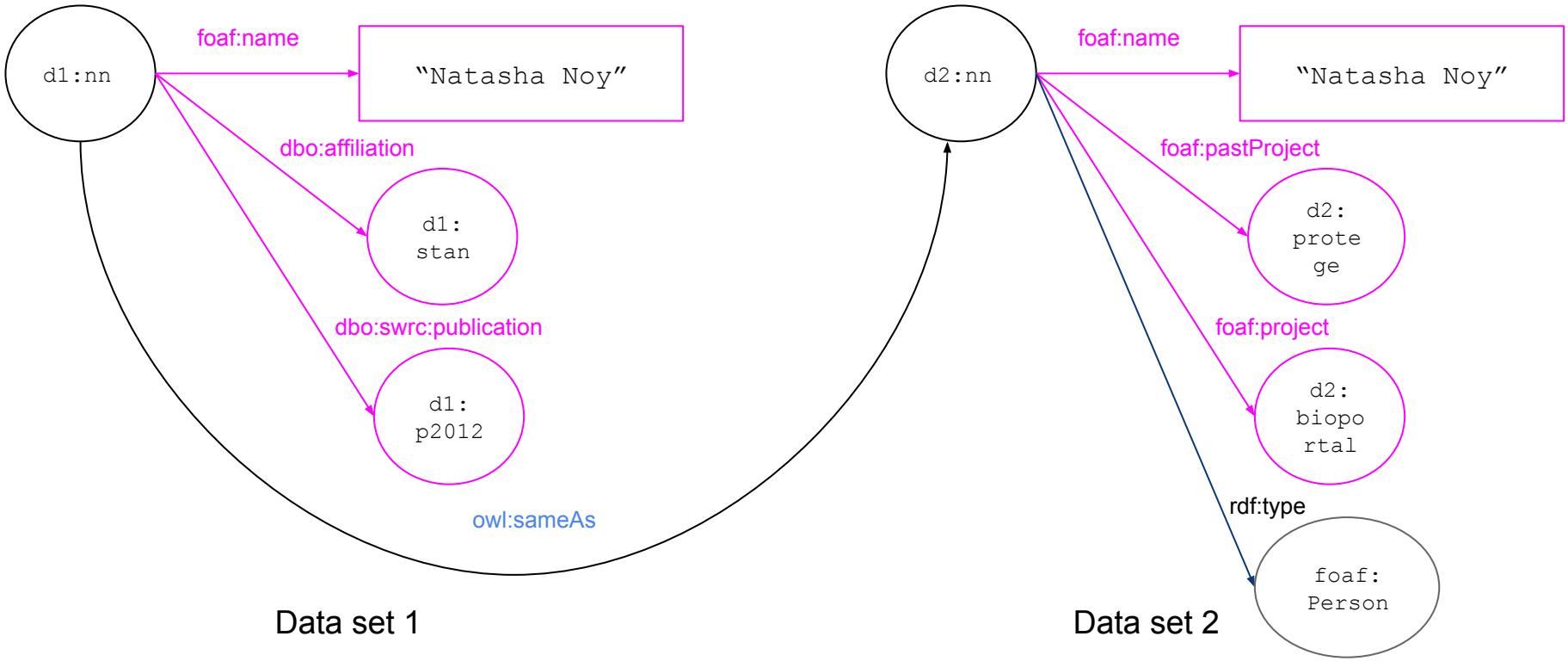


Part of the Koldfish framework

SeaStar Framework

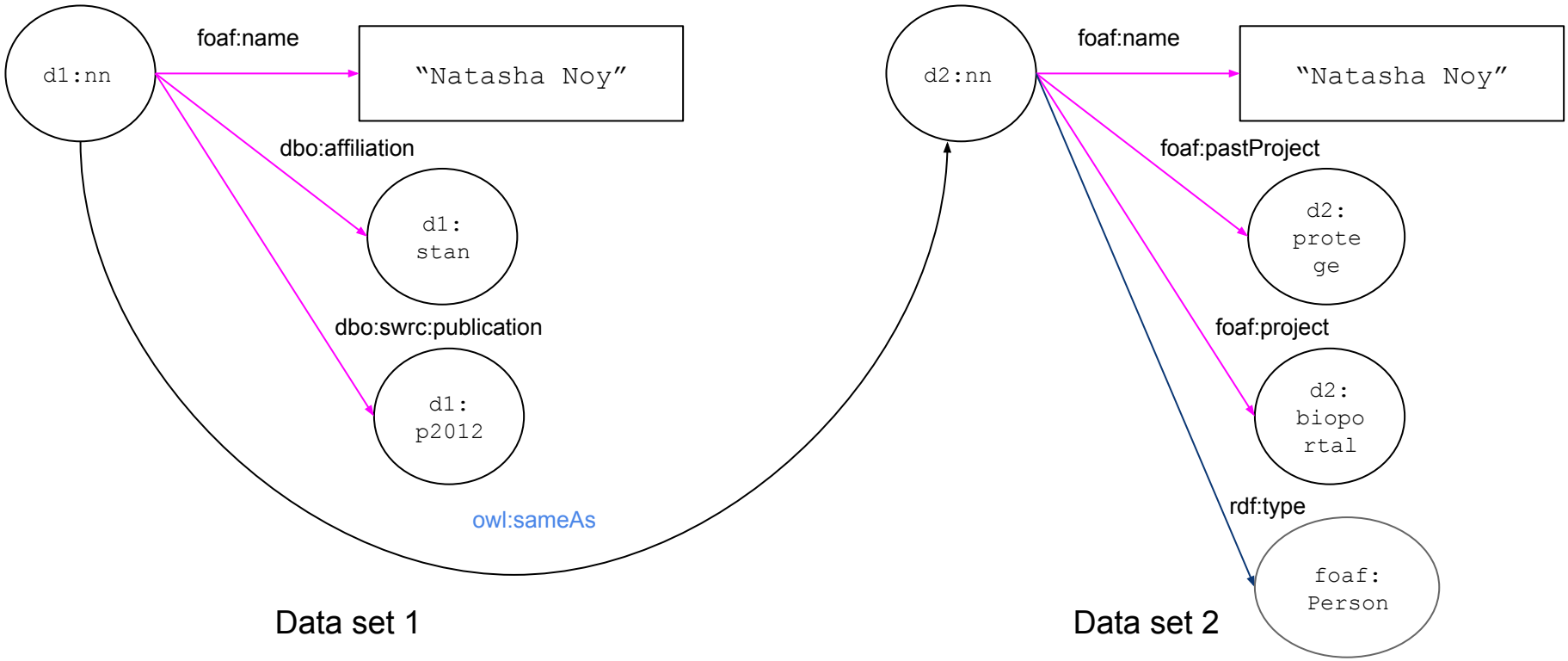
Our Solution

i) We create views of the data



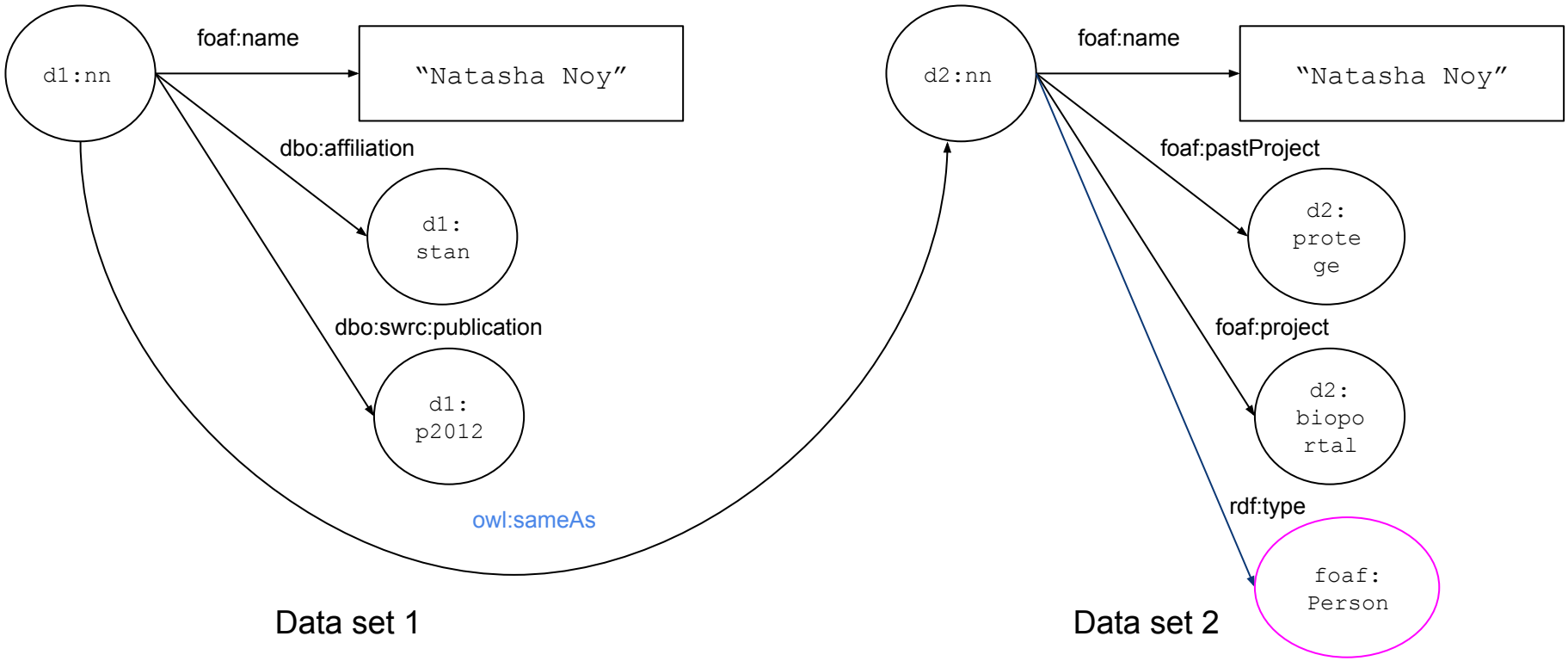
Our Solution

i) We create views of the data



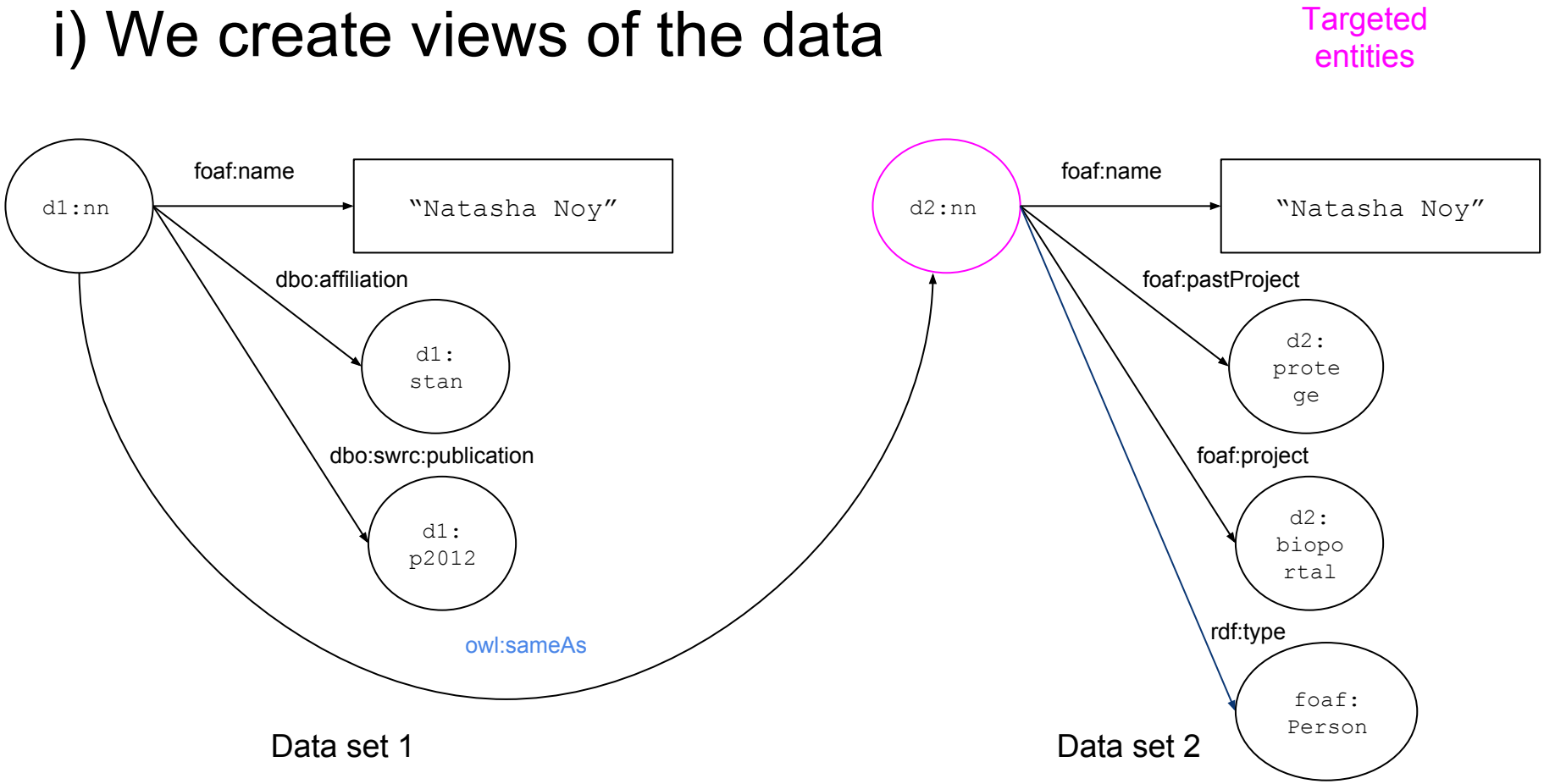
Our Solution

i) We create views of the data



Our Solution

i) We create views of the data



Our Solution

ii) Measure what each source entity gains for each view V, signaling redundancy.

Gain = diversity ($V_{\_with\_links}$) - diversity ($V_{\_without\_links}$)

Diversity as Shannon Entropy $H(S) = -\sum_{v \in V} prob(S = v) \times \log prob(S = v)$

E.g.:

descm(d1:nn,$D_{wo\ links}$) = {(foaf:name “Natasha Noy”), (o:likes d1:ISWC) }
descm(d1:nn,$D_{w\ links}$) = {(foaf:name “Natasha Noy”), (o:likes d1:ISWC), (foaf:name “Natasha Noy”) }

given DE, DE’

Gain = H(DE’) - H(DE) = - 0,082

Empirical Analysis

Methodology

- Data:
  - 35 data sets of the Linked Data Crawl 2014 [Schmachtenberg et al.,2014]
  - 5 topical domains (e.g. publications, gov)

Typelink	I	S	R	O	C	All
AEMET	0	0	96	0	57	153
BFS	1063	0	0	0	2862	3925
Bibbase	0	0	456	1401	0	1857
Bibsonomy	35646	0	2180	0	123080	160906
BNE	58	0	0	0	221	279
DNB	3577	0	8711	2278	55	14621
DWS Mannheim	71	0	296	39	926	1332
Eurostat	1182	0	2	0	1012	2196
Eye48	1	0	244	0	490	735
Fao	0	0	6	0	23	29
FigTrees	2	0	22	2	59	85
GeoVocab	11455	0	1759	113	7565	20892
GovWild	0	0	1998	0	0	1998
Harth	76	0	344	456	30	906
Icane	20	0	25	30	19	94
IMF	243	0	3	0	377	623
Korrekt	0	0	1174	0	7959	9133
L3S	1059	0	2478	1028	1089	5654

Typelink	I	S	R	O	C	All
LinkedGeoData	634	0	12	0	254	900
LOD2	26	0	282	50	180	538
NDLJP	1	0	178	60	267	506
Ontologi	0	0	5686	0	736	6422
Openei	6	0	323	0	203	532
Reegle	327	0	432	0	135	894
Revyu	1402	0	2145	1806	39772	45125
RodEionet	9	0	981	0	0	990
SemanticWeb	161	0	783	0	576295	577239
Sheffield	121	0	2189	1	27064	29375
Simia	6691	0	25113	0	38069	69873
Soton	50	0	352	0	160	562
SWCompany	2023	0	13473	421	43136	59053
TomHeath	7	0	34	4	6	51
Torrez	0	0	266	0	493	759
TWRPI	2	0	12	0	65	79
UKPostCodes	1	0	7	0	1	9

- Computed all measurements for all entities in the source data sets

See also:
https://goo.gl/6VVHBD

Methodology

- Measure Validation (following methodology by [Beckhamal et al., 2014])
  - Discriminative Measures
    - different values across data sets
  - Independent Measures
    - Spearman correlation

extended description
extended connectivity
increase number of vocabs

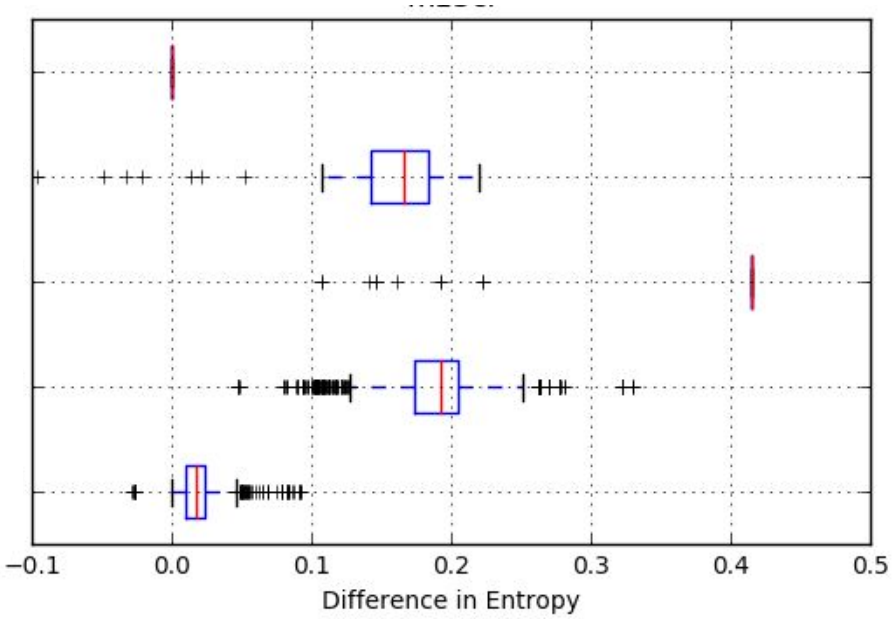
Measures	C m11	P,O m12	O m13	E m21	D m22	V m31
m11						
m12	1.00	0.29	0.58	0.55	0.55	
m13	1.00	-0.23	-0.22	0.76		
m21	1.0	0.96	0.04			
m22	1.0	0.02				
m31	1.0					

- Data Set Analysis (showcase)

See also: https://goo.gl/eD5kKX

Results

Gain in description - identity links
(analysing only properties)

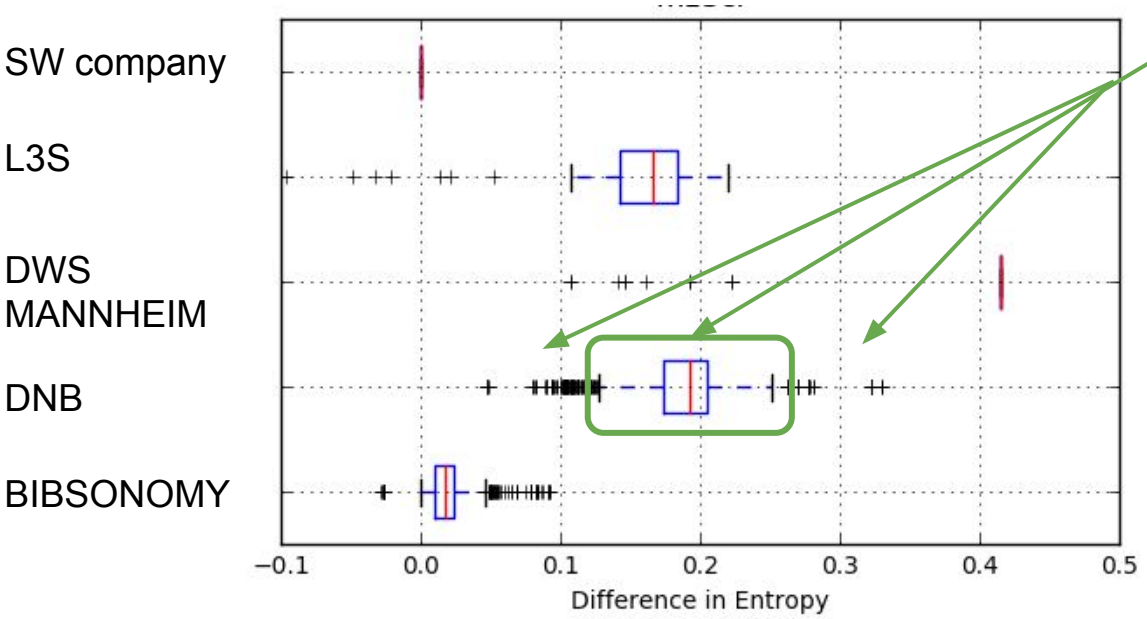


	#Entities	#Links	#AVG link p.entity
SW company	16	50	3,12
L3S	520	1059	2,04
DWS MANNHEIM	49	71	1,5
DNB	2782	3577	1,3
BIBSONOMY	19888	35646	1,8

Results

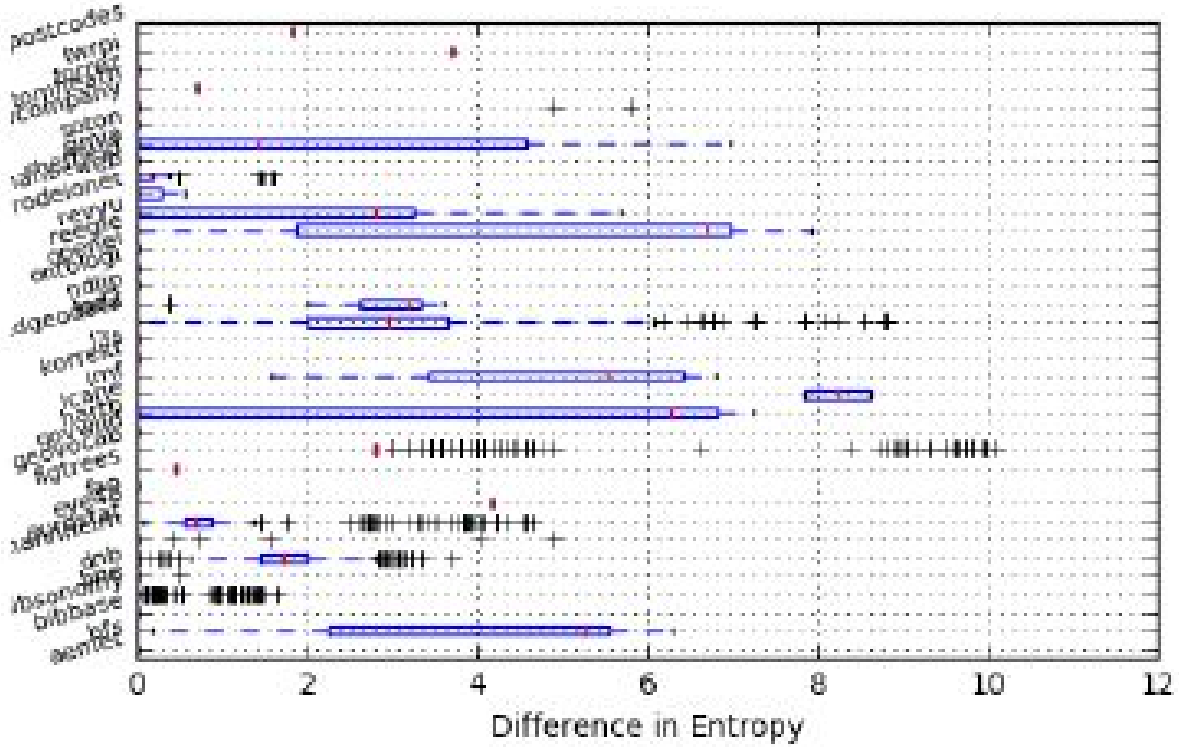
Gain in description - identity links
(analysing only properties)

Heterogeneity



Results

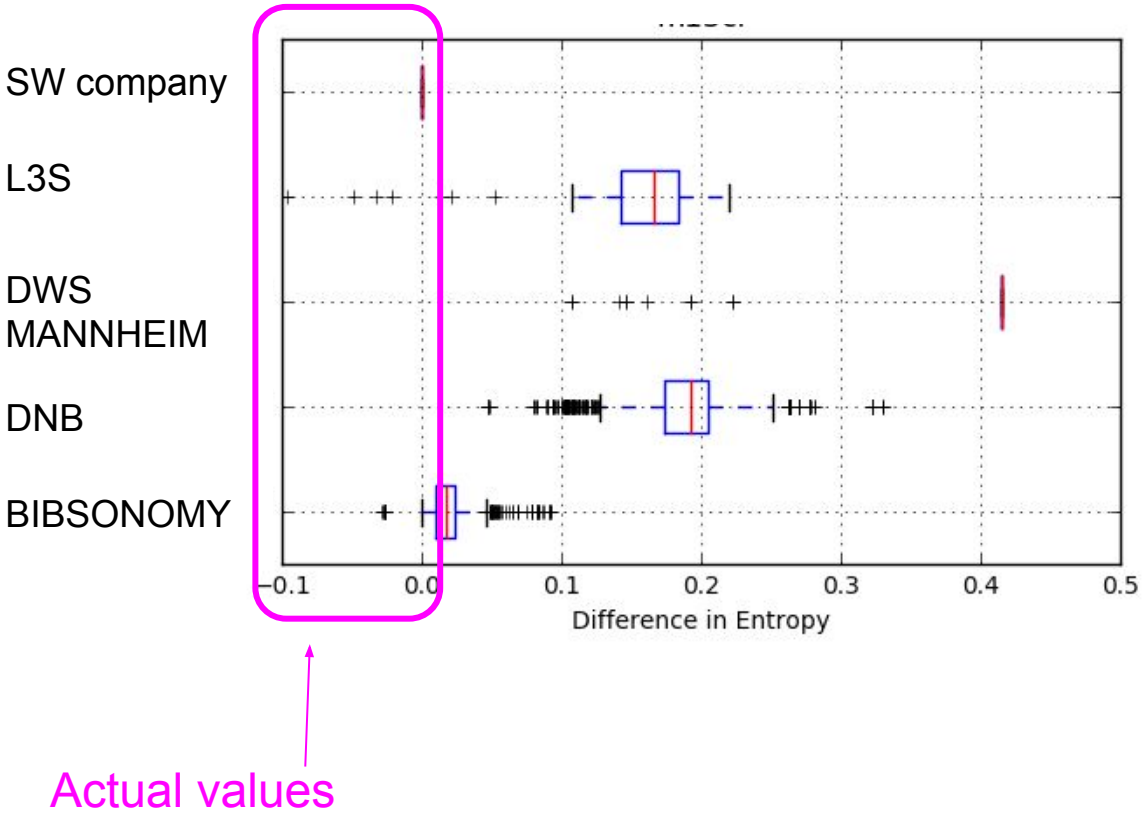
Gain in description - identity links
(analysing properties and values)



Extending the description (P1) is the principle with highest gain, and identity links the type of links with highest gain.

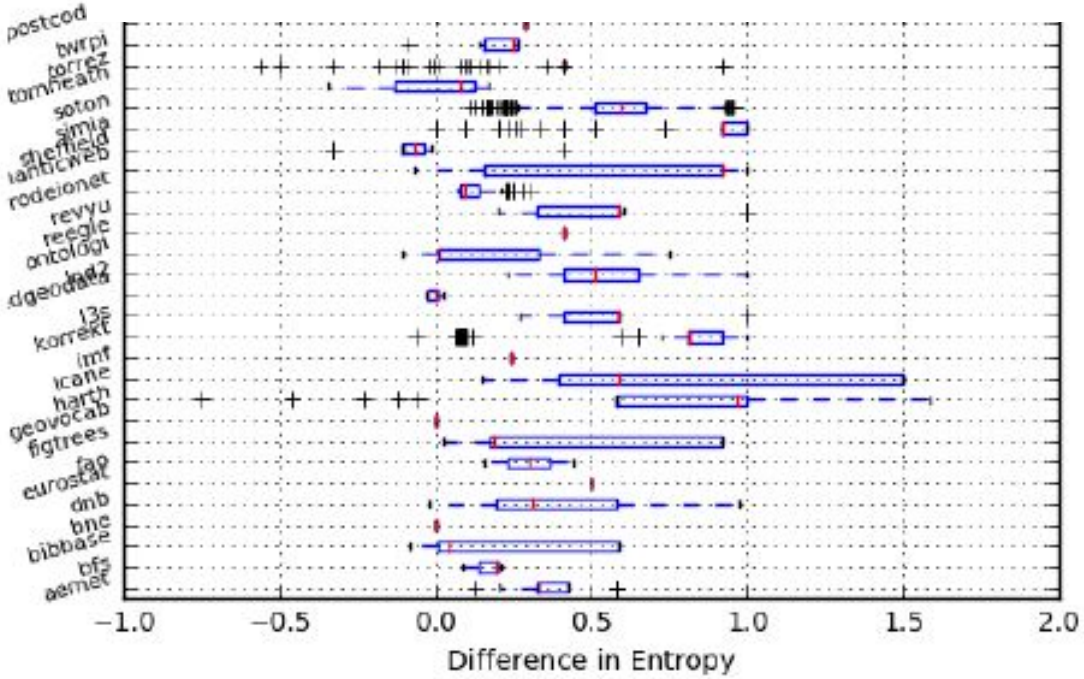
Results

Gain in description - identity links
(analysing only properties)



Results

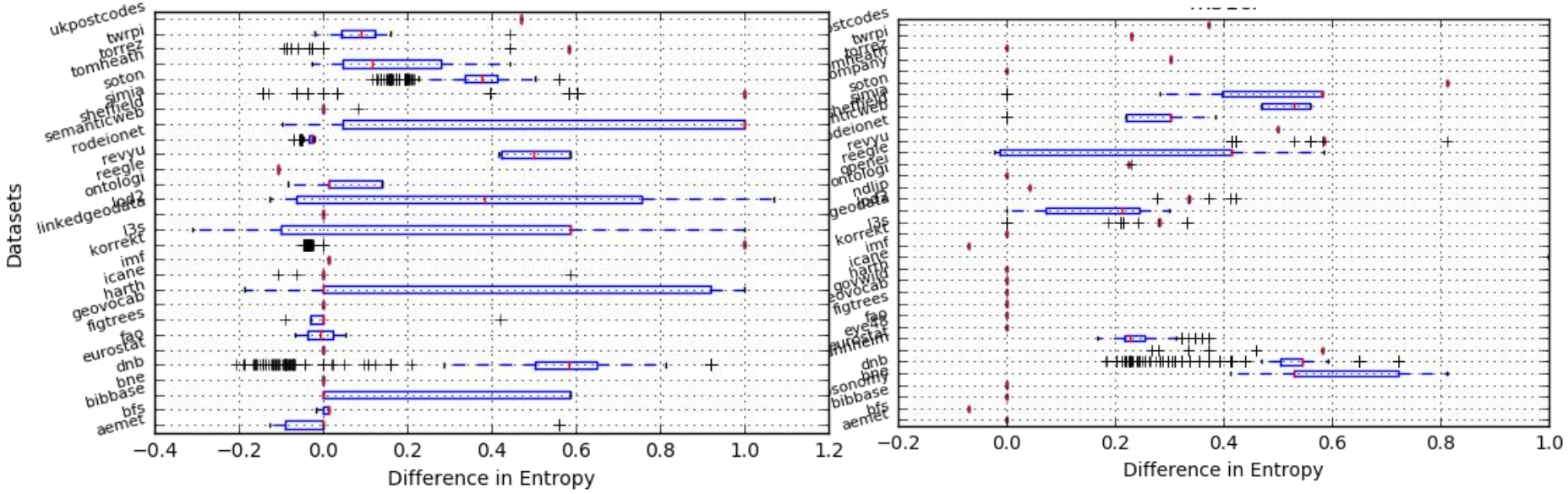
Gain in description - relationship links
(analysing only properties)



Redundancy appears when analysing only properties
(e.g. rdfs:seeAlso statements)

Results

Gain in # vocabularies
relationship and identity links



The number of vocabularies used only increases in a limited way.

Limitations and Future Work

- We do not include ontology mappings.
- An evaluation with data publishers should be done
  - relevance of each measurement
  - observing if (and how) they improve existing links based on them
- Would publishing the measurements as reports help data consumers in deciding what links to traverse?

Conclusions

- It is valuable to spot redundancy and counting new statements / new vocabularies is not sufficient.
- There is room for improvement
  - homogenizing what entities in a data set gain
  - enabling a higher gain
- If no match is found, other type of links can also enrich the entities in various dimensions.
- ***Main Contributions***:
  - Implemented and validated measures to assess the value (description, connectivity, classification) that links give to source entities, signaling redundancy.
  - Empirical analysis of 35 real LOD data sets.

Thank you!

csarasua@uni-koblenz.de

References

Berners-Lee, Tim. Linked Data: Design Issues. https://www.w3.org/DesignIssues/LinkedData.html

Max Schmachtenberg, Christian Bizer, and Heiko Paulheim (2014). Adoption of the Linked Data Best Practices in Different Topical Domains. In Proceedings of the 13th International Semantic Web Conference ISWC 2014.

Ermilov, I., Martin, M., Lehmann, J. and Auer, S., 2013, October. Linked open data statistics: Collection and exploitation. In International Conference on Knowledge Engineering and the Semantic Web (pp. 242-249). Springer Berlin Heidelberg.

Hu, W., Qiu, H. and Dumontier, M., 2015, October. Link Analysis of Life Science Linked Data. In International Semantic Web Conference (pp. 446-462). Springer International Publishing.

A Zaveri, A Rula, A Maurino, R Pietrobon, J Lehmann, S Auer. Quality assessment for linked data: A survey Semantic Web 7 (1), 63-93

References

Neto, C.B., Kontokostas, D., Hellmann, S., Müller, K. and Brümmer, M., 2016, April. Assessing quantity and quality of links between linked data datasets. In *Proceedings of the Workshop on Linked Data on the Web Co-located with the 25th International World Wide Web Conference (WWW 2016)*.

Christophe Guéret, Paul Groth, Claus Stadler, and Jens Lehmann. 2012. Assessing linked data mappings using network measures. In *Proceedings of the 9th international conference on The Semantic Web: research and applications* (ESWC'12).

Riccardo Albertoni and Asunción Gómez Pérez. 2013. Assessing linkset quality for complementing third-party datasets. In *Proceedings of the Joint EDBT/ICDT 2013 Workshops* (EDBT '13). ACM, New York, NY, USA, 52-59. DOI=http://dx.doi.org/10.1145/2457317.2457327

References

Shannon, C.E., Weaver, W. (1949) The Mathematical Theory of Communication, Univ of Illinois Press. ISBN 0-252-72548-4