Centre for
Data Analytics

Insight

Linked Data Profiling

Andrejs Abele

Supervisors: Paul Buitelaar, John McCrae

National University of Ireland, Galway

MixedEmotions

DCU

NUI Galway
OÉ Gaillimh

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

UCD
DUBLIN

sfi
Science Foundation Ireland

Bob

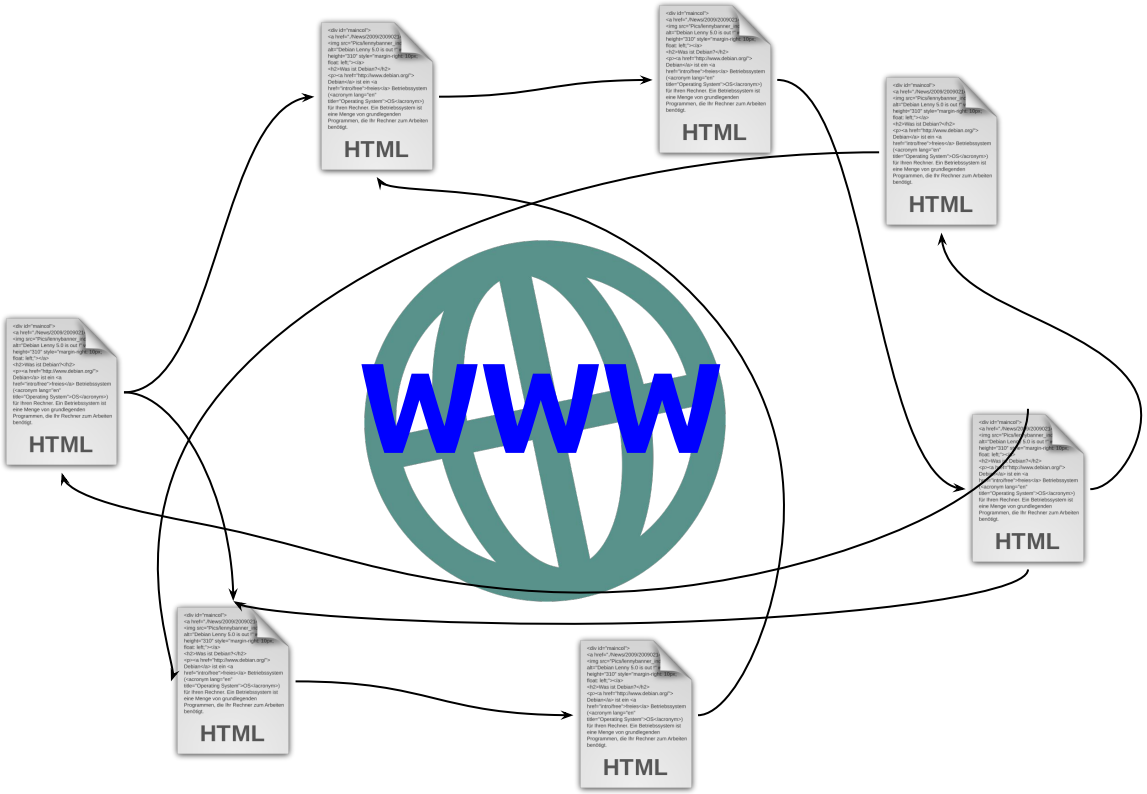
Works at government organisation

What is Linked Open Data?

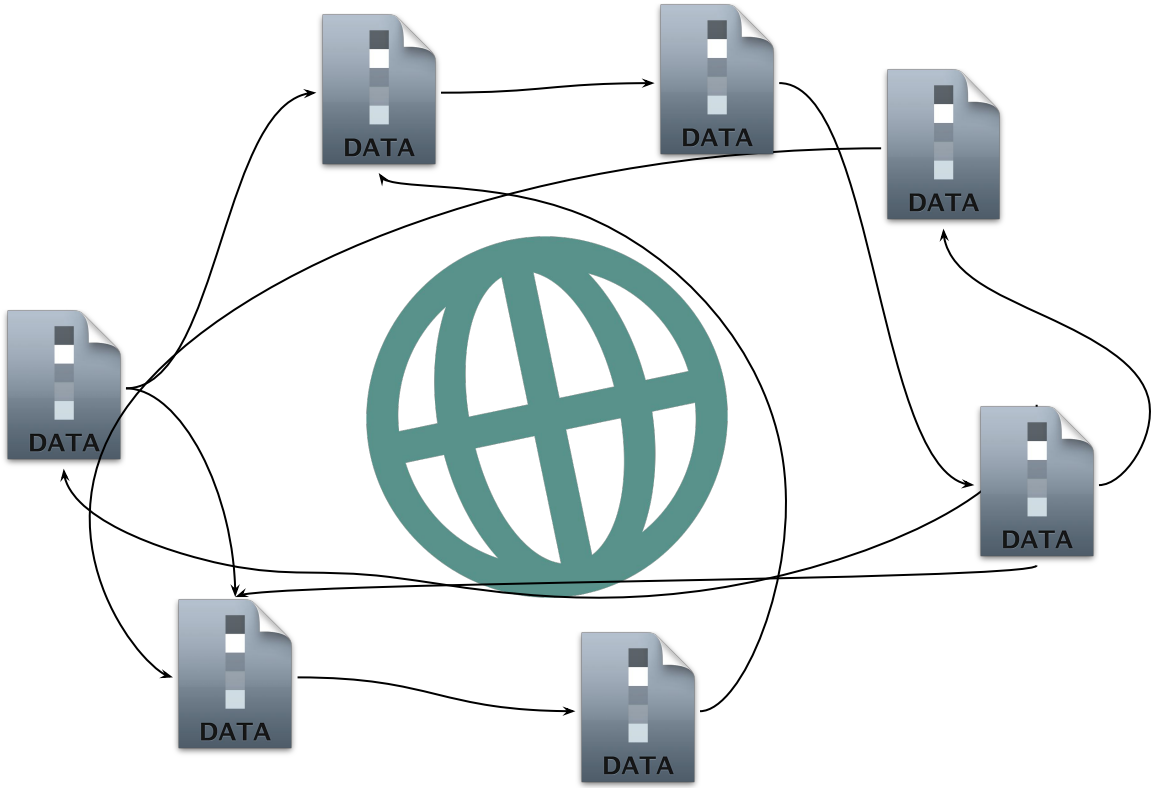
Bob's boss

Open Government Initiative
Linked Open Data

World Wide Web



World Wide Web of Data



Does not know what is inside the data
Does not know how to publish

Wants Bob to publish their agricultural datasets

Finds Datahub.io

Publishes dataset in Datahub.io

Tags the data with keyword "Government"

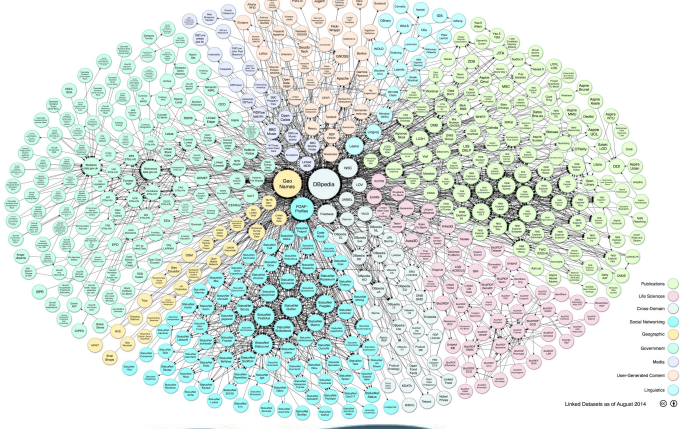
datahub

John

Is a developer

He wants to create an app that will help farmers to find best crops to plant on their land

He needs agricultural statistics for his system to work



He tries to find the data in LOD cloud

He knows that he is interested in agricultural data

LOD Cloud does not contain this type of category

He tries to find the data in Datahub

Because Bob tagged the dataset as "government", John can't find it

But John found other datasets connected to agriculture

datahub

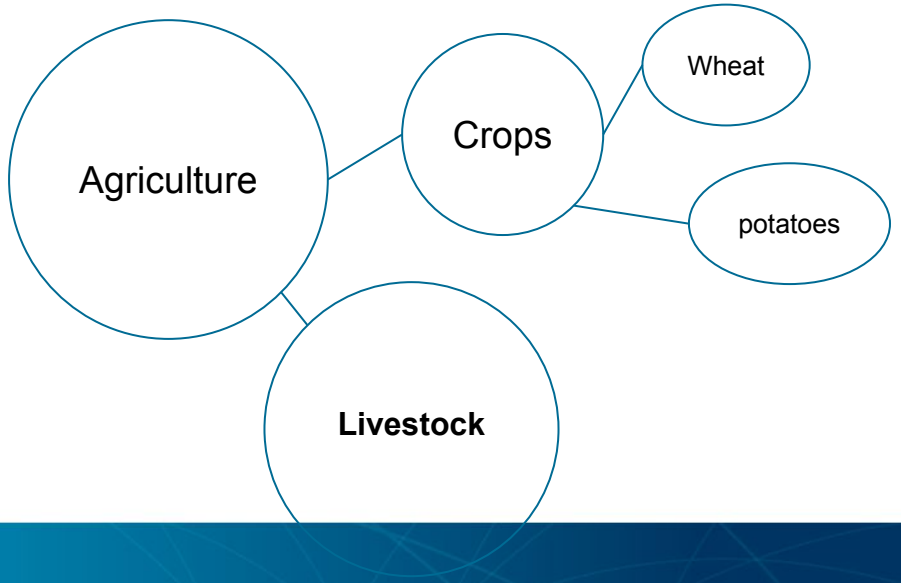
Because agriculture covers many subdomains, John manually had to go through the datasets to find the ones, related to crops

What if ...

Bob had a tool that would analyse his dataset and tell him:

- Your data is about crops
- which is subgroup of agriculture
- And it mentions locations, so you should link your data to "GeoNames" dataset

John had a place where he could browse the datasets based on different granularity



Problems and proposed solutions

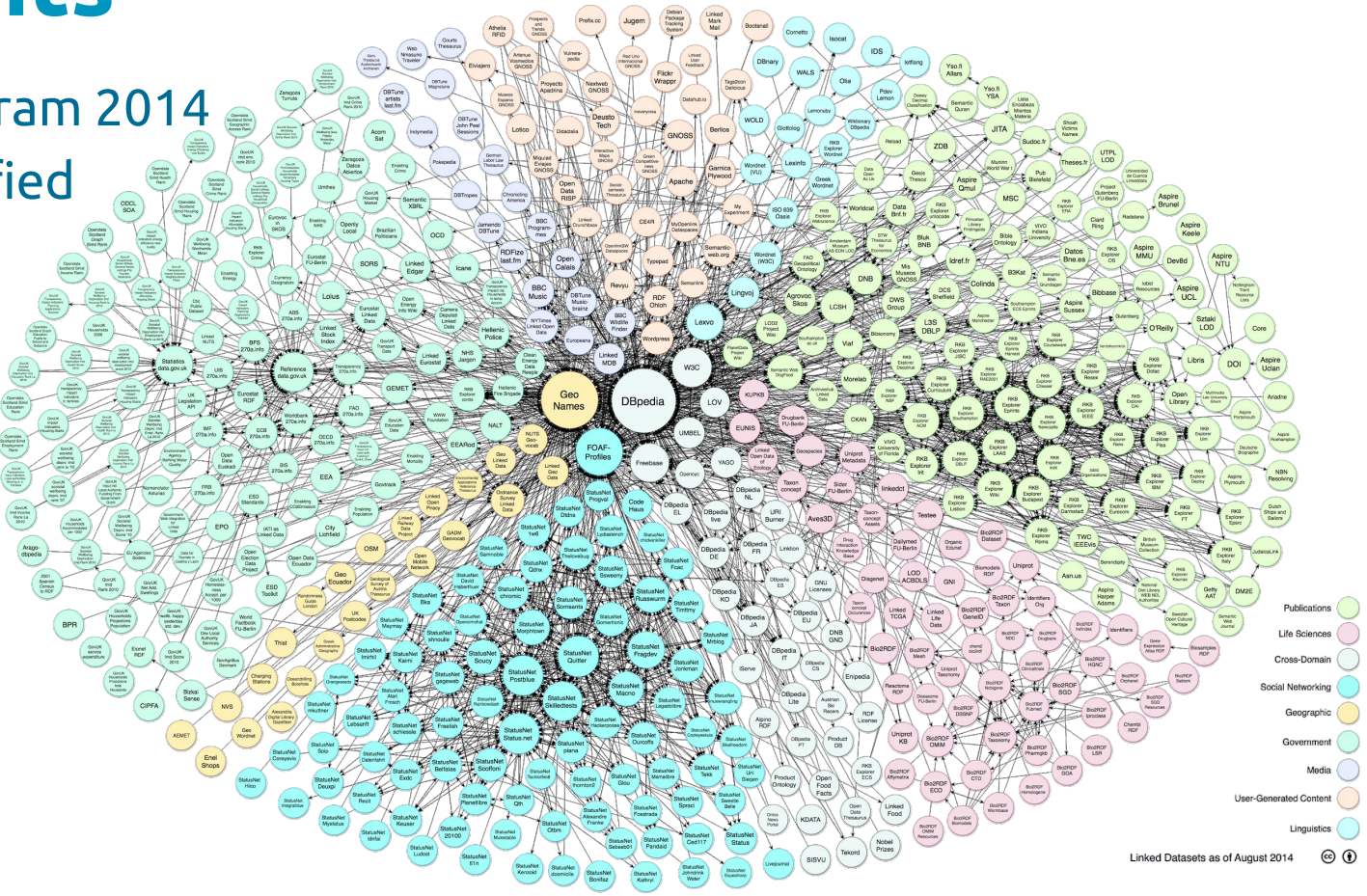
- There is no good overview of existing linked datasets
  Analyze the existing datasets and generate metadata for them
- Current LOD domain classification does not cover all the domains that are available
  Create hierarchical and more generic domain classification
- When people create LD datasets, they are not aware what exists and don't link their datasets to existing datasets
  Provide a tool that analyzes the dataset and provide list of related dataset
- Sometimes those who publish and annotate the datasets are not fully aware of the content
  Provide a tool that analyzes the dataset and automatically generates metadata about the dataset

Existing solutions for LD profiling

	Automatic domain identification	Statistics	Create VOID description	Visualisation	Uniqueness Analysis	Clustering and Labeling	Data browsing	Clean datasets	Pattern Analysis
LODStat		✔	✔						
RDF-stats		✔	✔						
Aether		✔	✔	✔					
Loupe		✔		✔			✔		
LOD Laundromat		✔	✔	✔				✔	
ProLOD++		✔		✔	✔	✔	✔		✔
Our Approach	✔	✔	✔	✔	✔	✔			

Experiments

- LOD cloud diagram 2014
- Manually classified



- Automatically classify datasets using SVM classifier

Experiments

1. Train classifier using Support Vector Machine:
    a. using datasets classes and properties as features
    b. using datasets classes and properties as features and enrich with tags from Linked Open Vocabularies
    c. using tags from DataHub as features

Datasets (405)

Domain	Datasets	Largest*	Smallest*	Average*	Popular dataset
Life_sciences	35	13,081,788,247	11,091	766,662,173	BioPortal
Geography	29	3,000,000,000	30,583	218,860,608	GeoNames
Cross_domain	25	1,200,000,000	1,975	109,693,105	DBpedia
Government	65	8,000,000,000	500	66,885,645	reference.data.gov.uk
Publications	111	1,000,000,000	3,500	38,912,951	DBLP Computer Science Bibliography (RKBExplorer)
Media	13	250,000,000	23,861	23,997,770	New York Times
User_generated	52	416,732,232	4,866	18,488,968	Linked Crunchbase
Linguistics	34	32,916,476	4,374	4,064,094	BabelNet
Social_networking	41	4,050	3	504	StatusNet

* number of N-triples

Classification by classes and properties

1. Extract URIs of properties and classes from dataset
   a. Classes = all subjects that have predicate "rdf:type" and object "owl:Class"
   b. Properties = all predicates
2. Transform classes and properties into binary feature vectors
3. Train Support Vector Machine classifier using LOD cloud dataset
4. Cross-validation using precision and recall as metrics

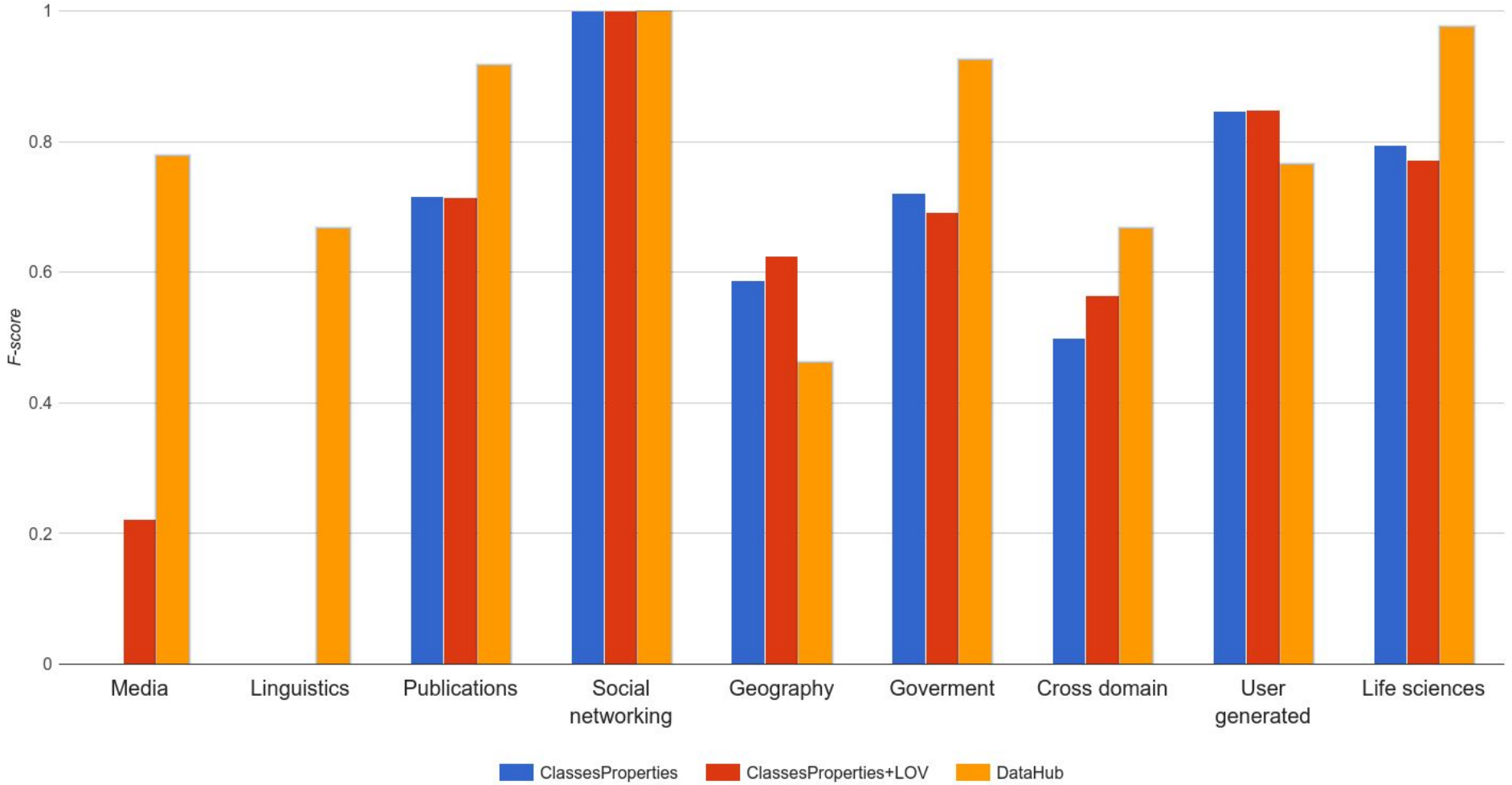
Classification by classes, properties and Linked Open Vocabulary tags

1. Extract URIs of properties and classes from dataset
   a. Classes = all subjects that have predicate "rdf:type" and object "owl:Class"
   b. Properties = all predicates
2. Link classes and properties to LOV tags
3. Transform classes, properties and LOV tags into binary feature vectors
4. Train Support Vector Machine classifier using LOD cloud dataset
5. Cross-validation using precision and recall as metrics

Classification by Datahub tags

1. Extract tags describing datasets from DataHub
2. Transform tags into binary feature vectors
3. Train Support Vector Machine classifier using LOD cloud dataset
4. Use Precision and Recall as metrics
5. Use cross-validation to evaluate the classifier

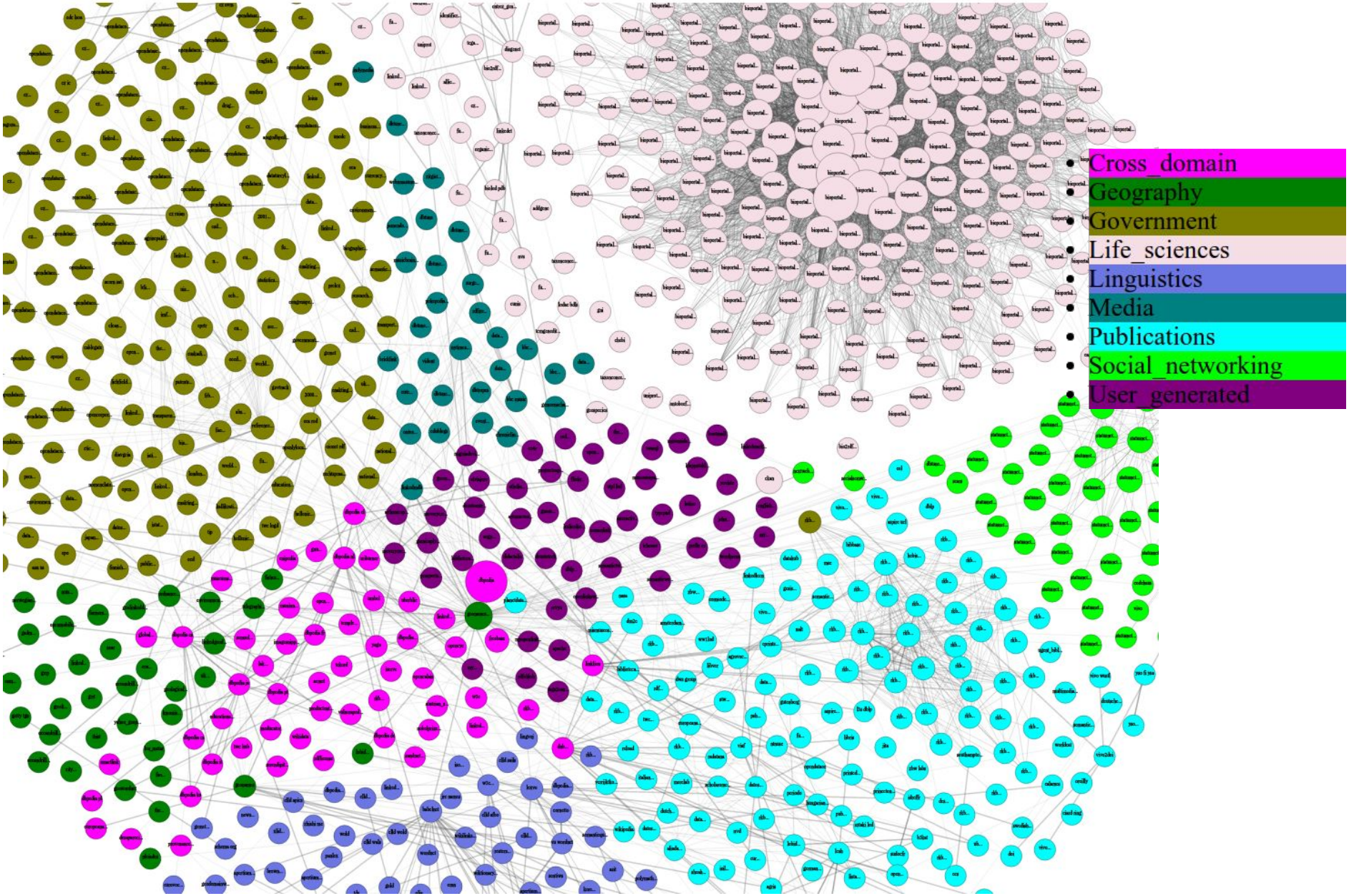
F-Measure for different domains using SVM



DataHub > ClassesProperties , p=0,01
DataHub > ClassesProperties + LOV , p= 0,01
ClassesProperties > ClassesProperties + LOV , not significant

Further work

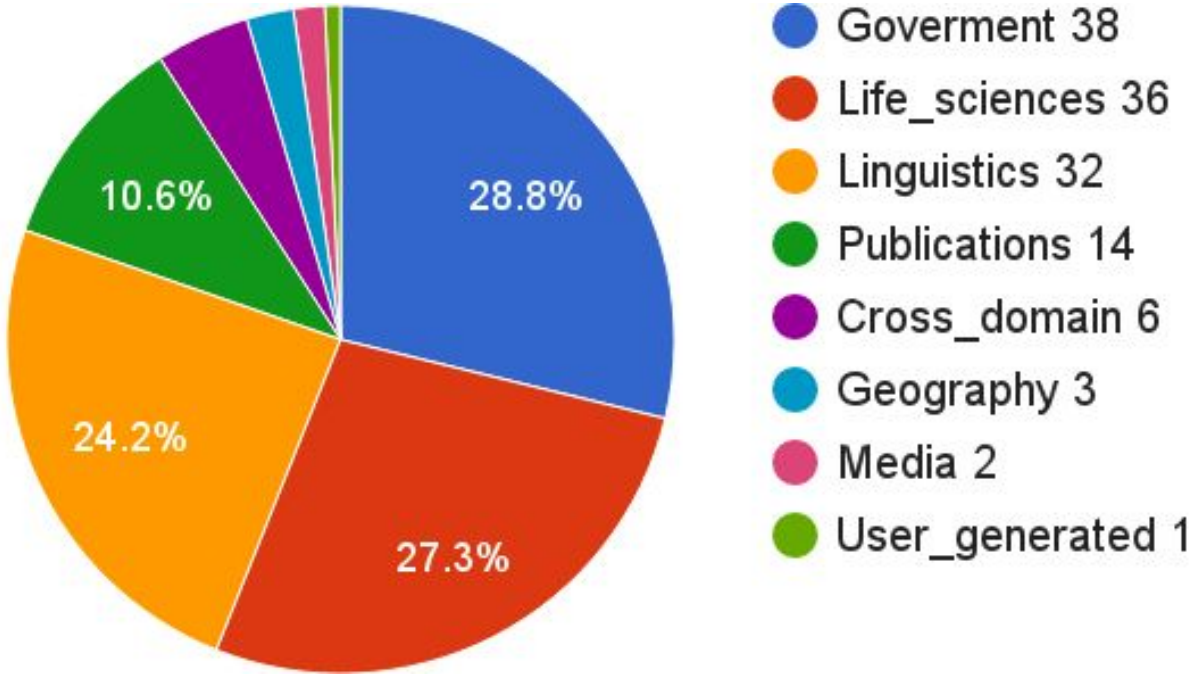
- Identify more representative domain categories for linked data classification
- Identify approach for creating category hierarchy
- Create gold standard/ test dataset
- Identify best visualisation to represent the metadata from the analytics


Cross_domain
Geography
Government
Life_sciences
Linguistics
Media
Publications
Social_networking
User_generated

Additional slides

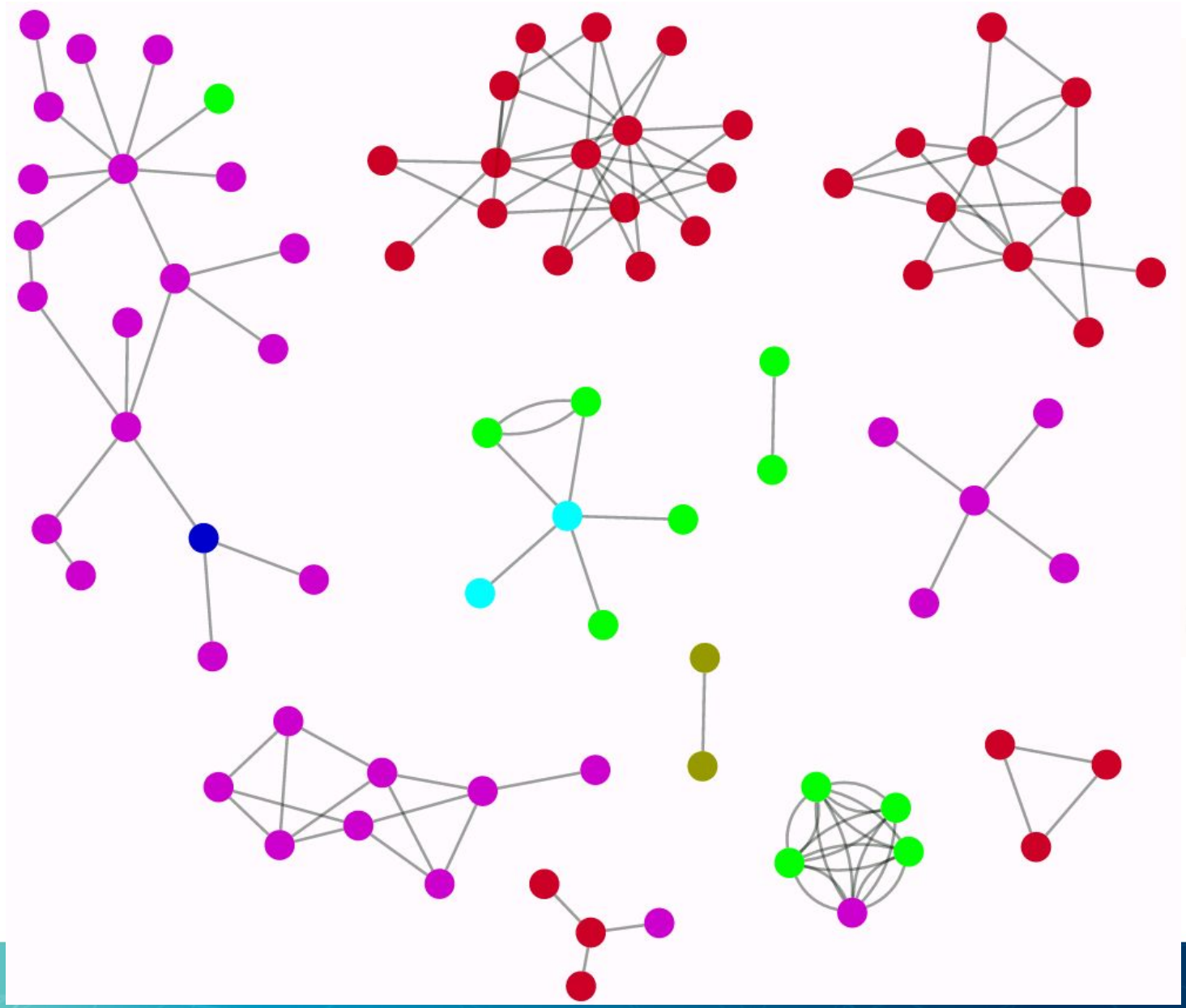
Datasets

- Gold standard : 132 datasets from LOD cloud

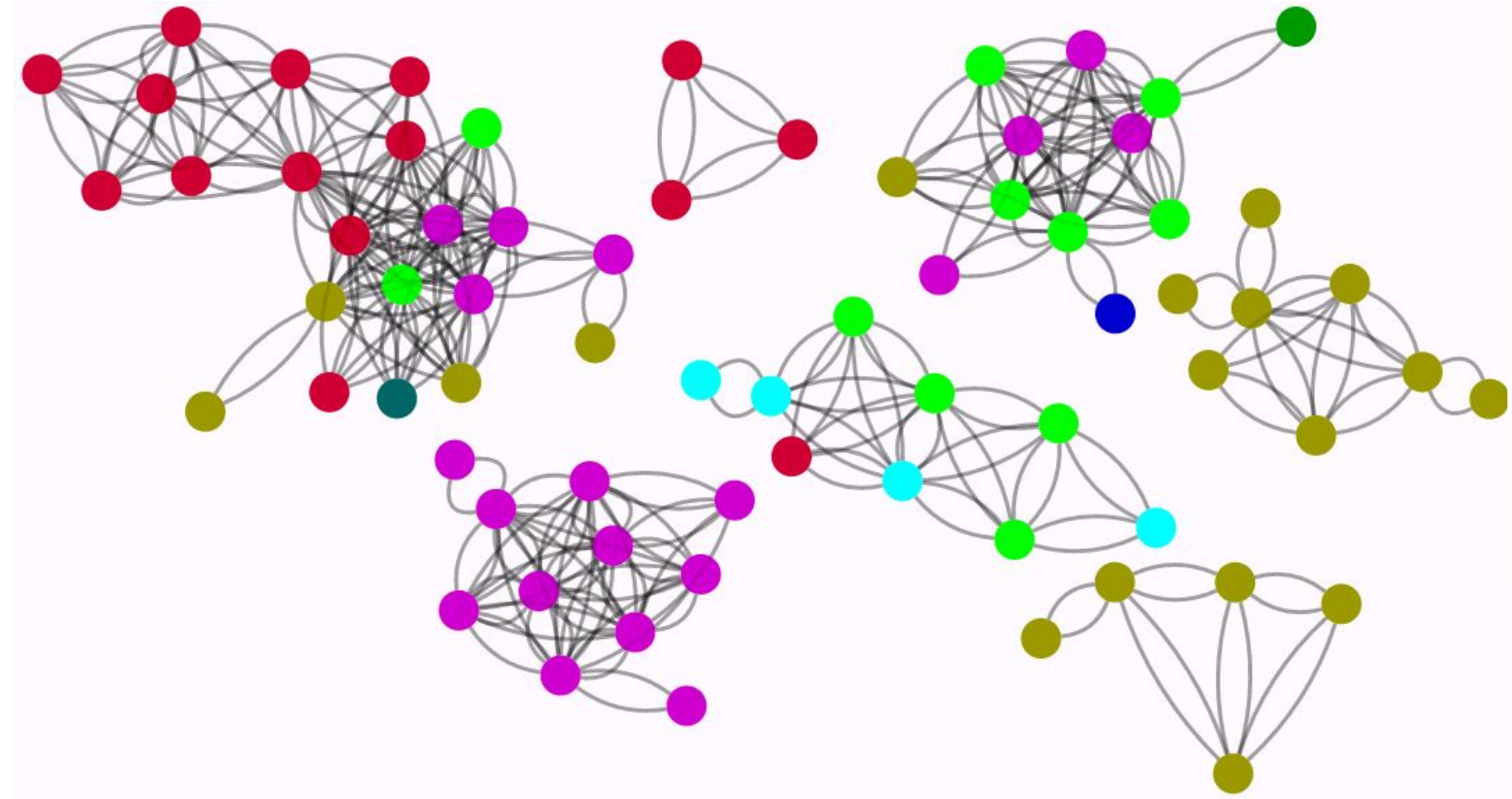


- Methods for dataset collection :
  - Links between datasets (from datahub)
  - Extract terms from literals in datasets

Clustering based on links



Clustering based on text similarity (1≤Jacc≥0.5)



Cross_domain
Geography
Goverment
Life_sciences
Linguistics
Media
Publications
User_generated

Results based on B-cubed measure

	Precision	Recall	Fscore
Based on links	0.887301587302	0.384312512884	**0.536327967191**
1≤Jacc<0	0.413362189972	0.618381496671	0.49550199921
1≤Jacc≥0.5	0.593073593074	0.482368041315	0.532022823553