INF 558 Project Summary:

Marvel and DC Comic Books Knowledge Graph

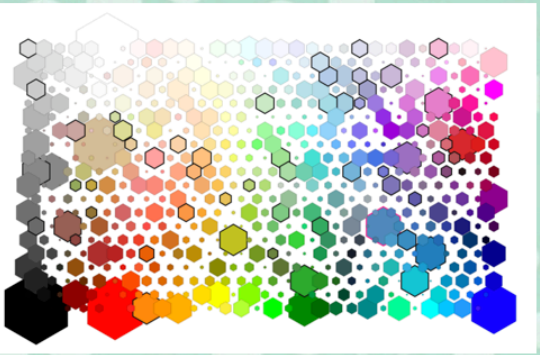
Basel Shbita, Ke Xu

11/30/18

Intro (Reminder)

first-appearance-date
gender
publishers
birth-date
death-date
origins
powers
relationships
photos

analysis



publication-date
characters involve
cast data
reviews/score

publication-date
setting
characters involved
cast data
reviews/score

Goals (Reminder)

Discover
New and interesting facts!

Search
More easily!

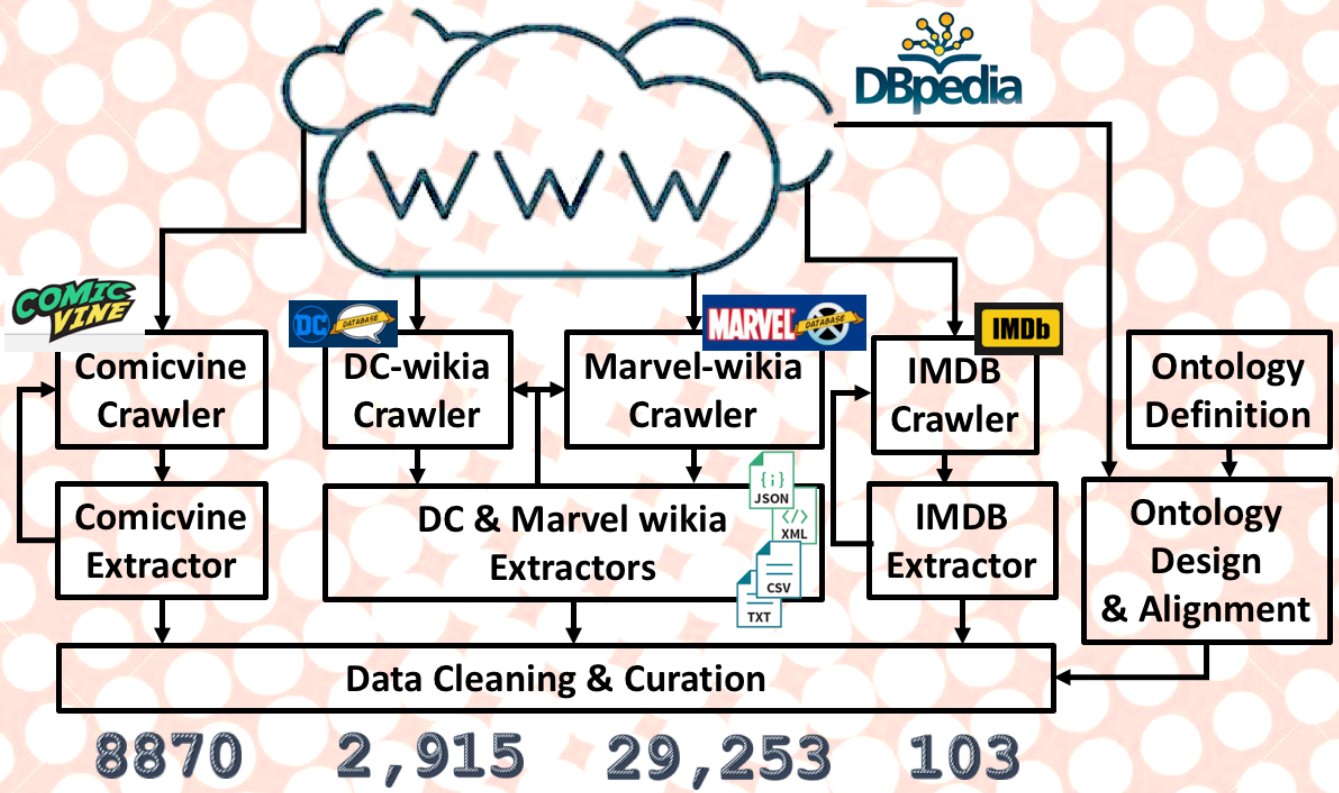
Expand
Their knowledge & read issues and/or watch movies they were not aware of!

Explore
Further in their favorite domain

Analyze
Data!
Characters, movies & comic-book issues

Capture
Relations Between comic-book characters & associated colors!

The Big Picture - Part 1!



Crawling Challenges

Problem

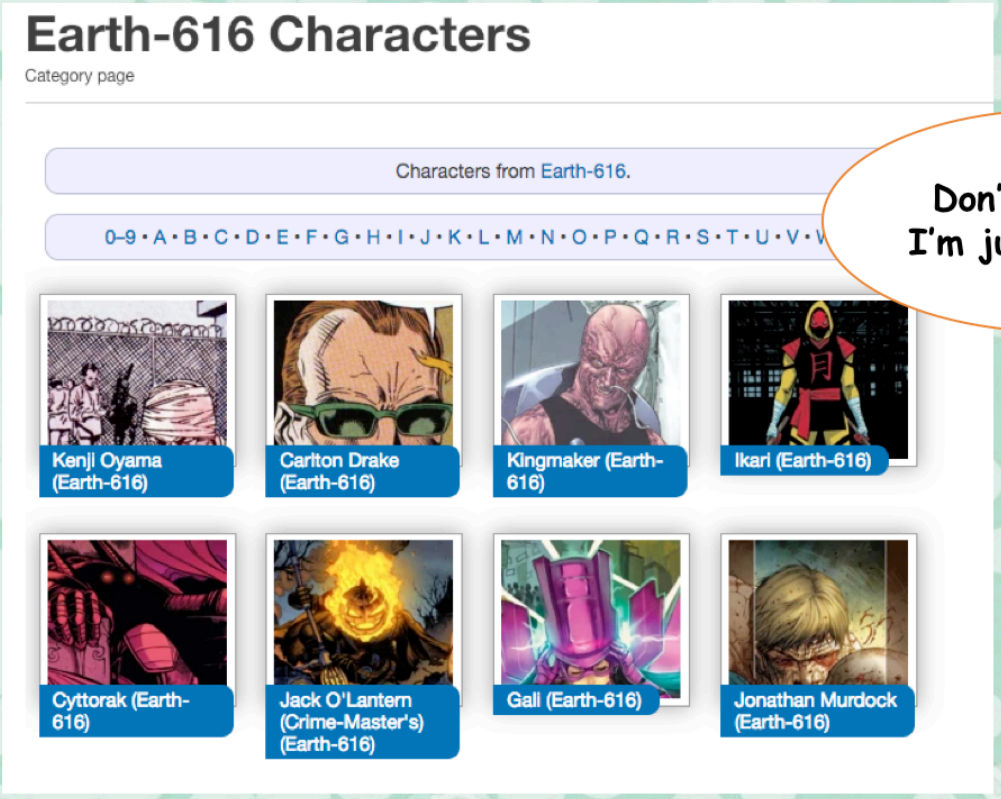
SCALE (Ms of pages)

Solution

Focused approach

Generate hub pages
to easily reach
authority pages

Don't mind me!
I'm just crawling!

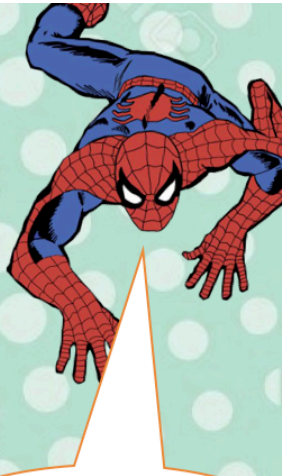


Crawling Challenges

Problem

Irrelevant Pages

I.E. Comicvine 1st crawl result:
Crawled: 68,294
Usable: 10,902



Don't mind me!
I'm STILL crawling!

Solution

Pattern Matching

Implement strict pattern matching and re-crawl some of the data.

Extraction Challenges

Problem

Structure Inconsistency

Solution

Rule-Based System

Implement rules to segment data and look for syntactic patterns, regex & structure features

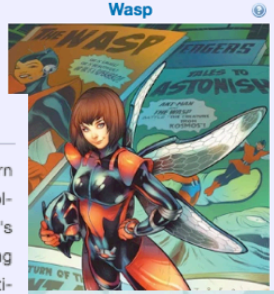
Nadia Pym (Earth-616)

In accordance with the naming conventions, it has been suggested that this page be moved to:
Nadia Van Dyne (Earth-616) (Discuss).
Please do not move this page yet, as the correct name could still be in discussion and may change again soon. A robot will make the necessary page moves once any discussions are resolved.

MARVEL

Contents [show]

Wasp



Origin

...o was born into captivity and followed in her father's footsteps by mastering the Pym Particle, ultimately escaping to become the new Wasp.

Universe	Earth-616
Creators	Mark Waid, Alex Ross

First Appearance

***Free Comic Book Day* Vol 2016 Civil War #II**
(May, 2016)

Roger Norvell (Earth-616)

Contents [show]

History Edit

Roger Norvell, or Red as he prefers to be known, joined Harris Hobbs' film crew along with Joey Burnett to document the first TY show special of the real-life Norse

Red Norvell

Origin

Origin	Human
Universe	Earth-616
Place of Birth	Somewhere in the USA
Creators	Roy Thomas, John Buscema

First Appearance	Last Appearance
***Thor* #273** (July, 1978) (Named) ***Thor* #276** (October, 1978) (As Thor)	***Journey into Mystery* #513** (October, 1997)

Data Cleaning Challenges

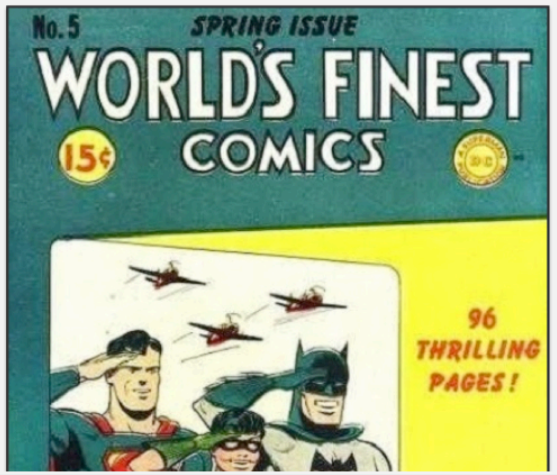
Problem

Format Inconsistency

Issue details

Name	Important Women in His Life
Volume	Batman
Issue Number	208
Cover Date	February 1969
In Store Date	November 12, 1968

Name	Jungle Jeopardy
Volume	Batman
Issue Number	209
Cover Date	February 1, 1969



Volume	World's Finest Comics
Issue Number	5
Cover Date	Q1 1942
In Store Date	February 13, 1942

Solution

Classify & Transform

Classification determines the **Transformation** type

Data Cleaning Challenges

Problem

Multiple Values Same Meaning

```
SELECT ?pred (count(?pred) as ?count)
WHERE {
  ?pred rdfs:subPropertyOf mdcu:relative ;
}
GROUP BY ?pred
ORDER BY DESC(?count)
```

1,802 entries !!!

	var
1	"single"
2	"married"
3	"widowed"
4	"divorced"
5	"separated"
6	"engaged"
7	"widower"
8	"widow"

mdcu:relative_ex_nephew_by_marriage_and_ex_adoptive_son_in_law

mdcu:relative_ex_nephew_in_law

mdcu:relative_ex_niece_by_marriage_and_ex_adoptive_daughter_in_law

mdcu:relative_paternal_great_great_great_grandfather

mdcu:relative_paternal_great_great_great_grandmother

mdcu:relative_paternal_great_great_great_great_grandfather

Solution

Clustering

Cluster the different types
Define a 'fixed' set
Re-classify and transform

Ontology Challenges

Problem

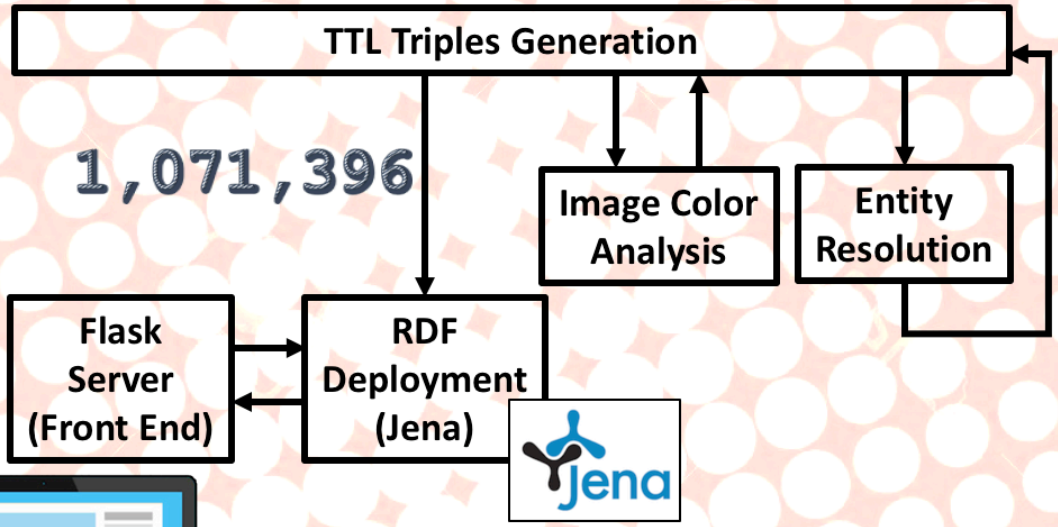
Capture concepts With minimal effort

Solution

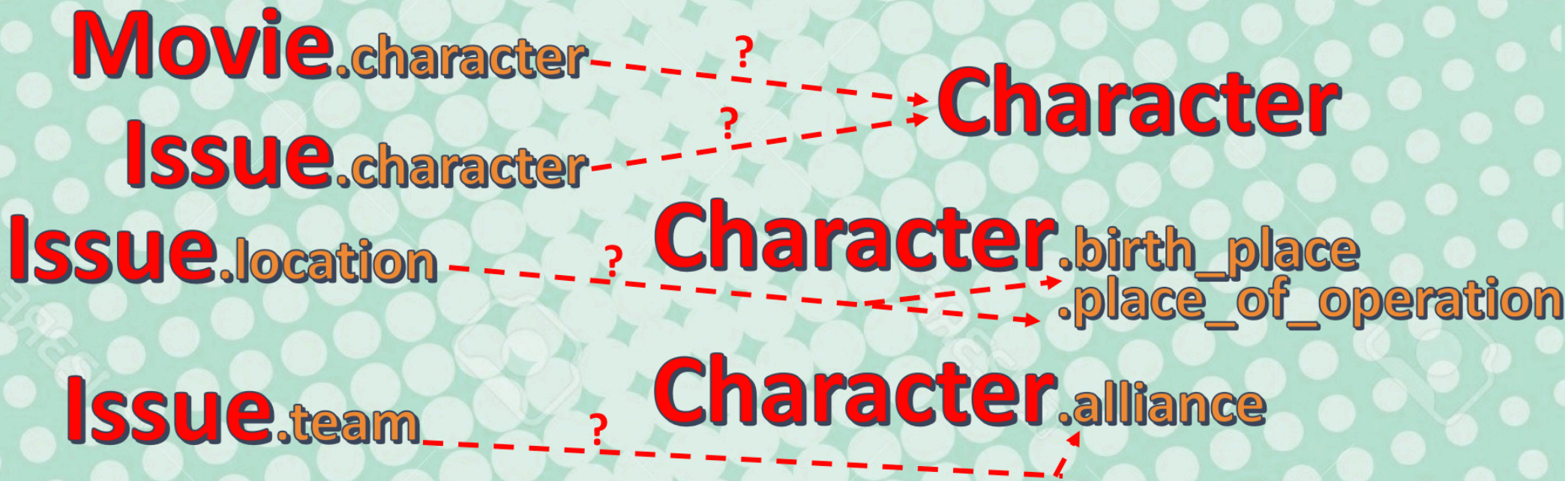
Combine pre-defined ontologies with custom one

We utilized existing ontologies as schema.org, xsd, foaf and dbpedia (dbp, dbo, dbr)

The Big Picture - Part 2!



Entity Resolution Problems



Solution

ISI-RLTK

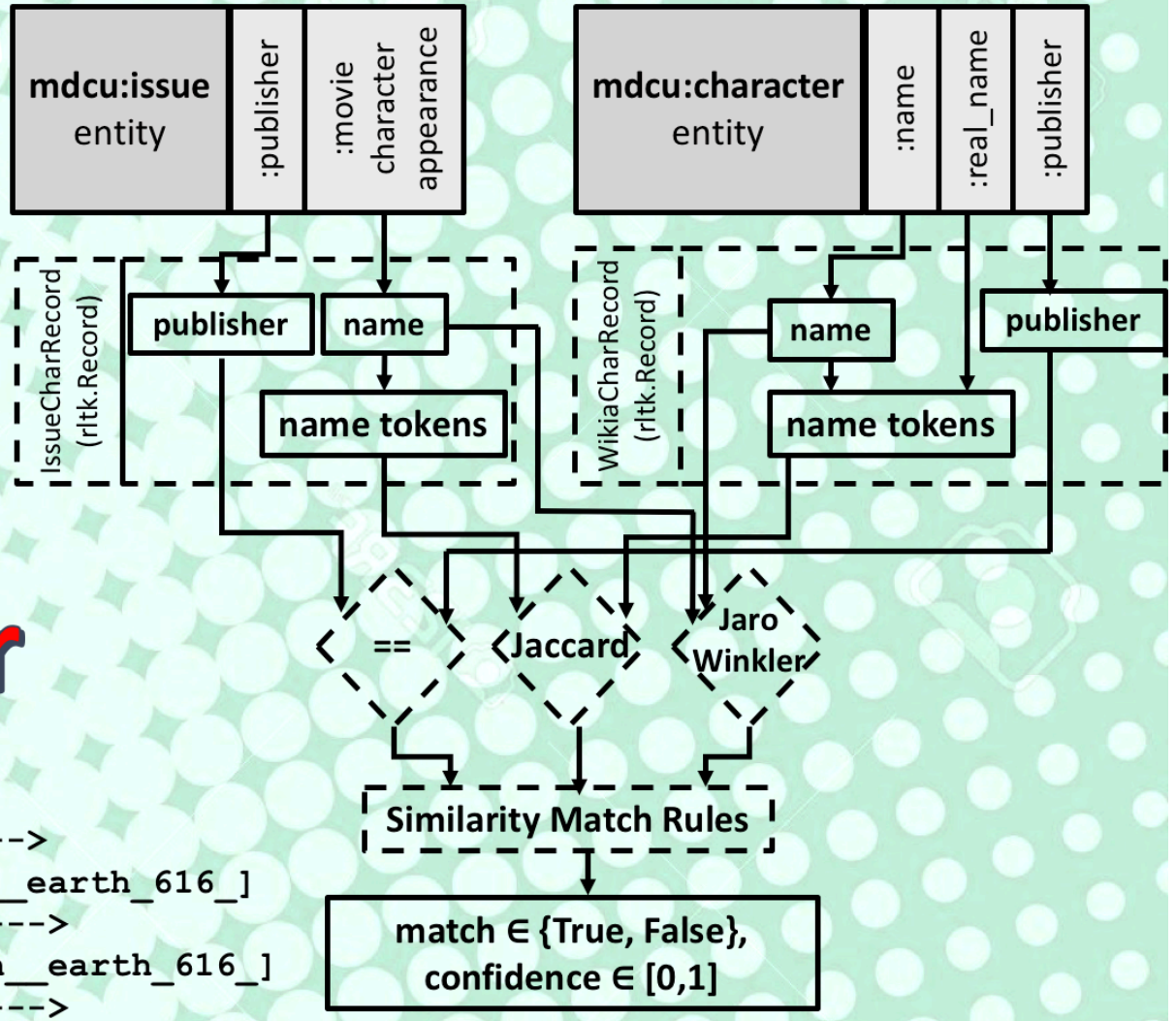
Use different attributes we have for each entity to perform entity resolution

Entity Resolution

EXAMPLE

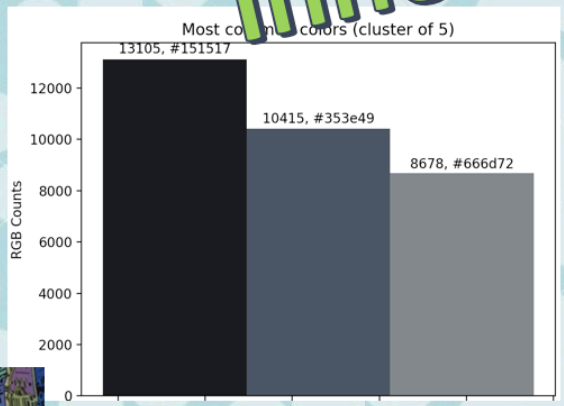
Issue.character

Character



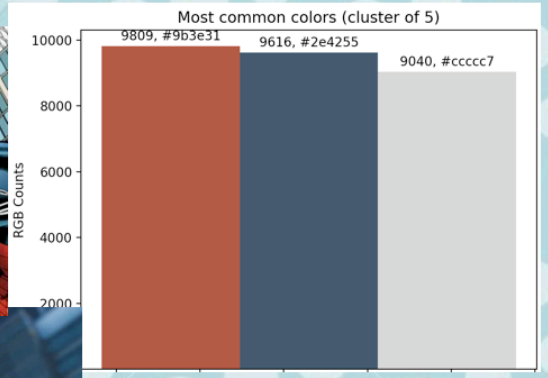

```
[mdcu:_morgan-stark_4005-19688_] ---96.19%--->
                    [mdcu:_wiki_morgan_stark__earth_616_]
[mdcu:_heather-glenn_4005-10365_] ---99.17%--->
                    [mdcu:_wiki_heather_glenn__earth_616_]
[mdcu:_richard-nixon_4005-13426_] ---77.50%--->
                    [mdcu:_wiki_richard_nixon__new_earth_]
```

Color Analysis Innovation



['#151517', '#353e49', '#666d72']

[(('black', (0, 0, 0)), 23520)]



['#9b3e31', '#2e4255', '#ccccc7']

[(('red', (255, 0, 0)), 9809),
(('blue', (0, 0, 255)), 9616)]

For each character: (>30,000)

download image (given img-url) → resize → for cluster in range(2, 7):

codes = scipy.cluster.vq.kmeans(..)

counts = scipy.histogram(..)

classify_by_euclidean_distance(codes, counts)

→ extract dominant colors (> 70%) → plug-in to graph (rdflib)

User Interface

Flask
web development, one drop at a time

Jena

Bootstrap

main page

dropdown menu for class selection

predicates are generated automatically in BE

responsive pie and bar charts generated automatically

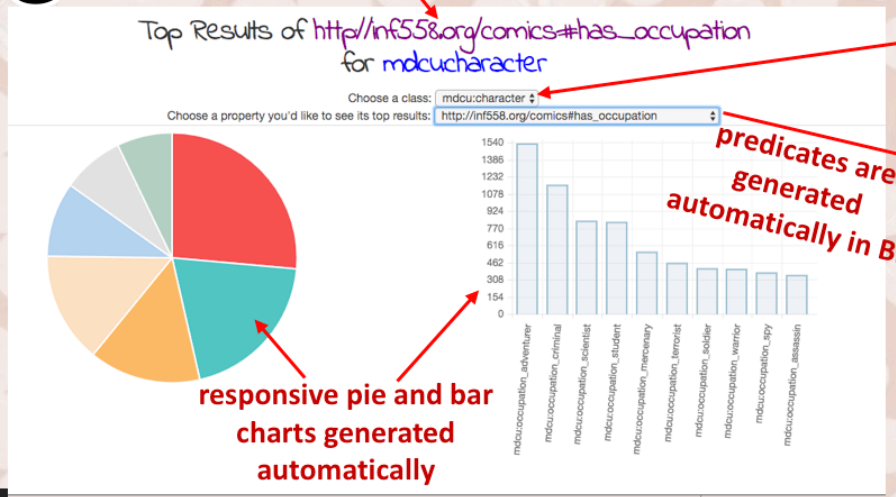
links to a description page

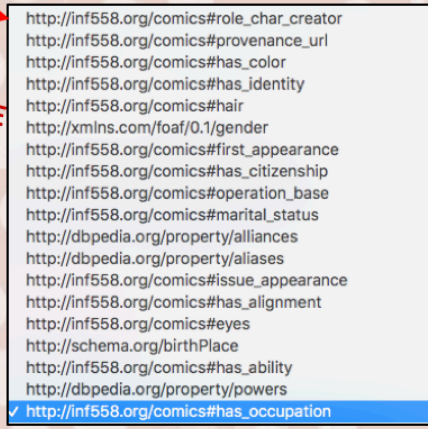
perform text search within results

sort by s,p,o

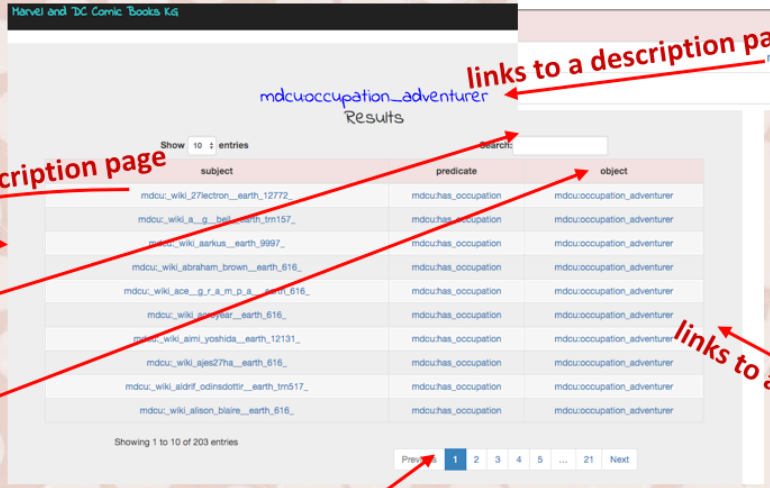
pagination

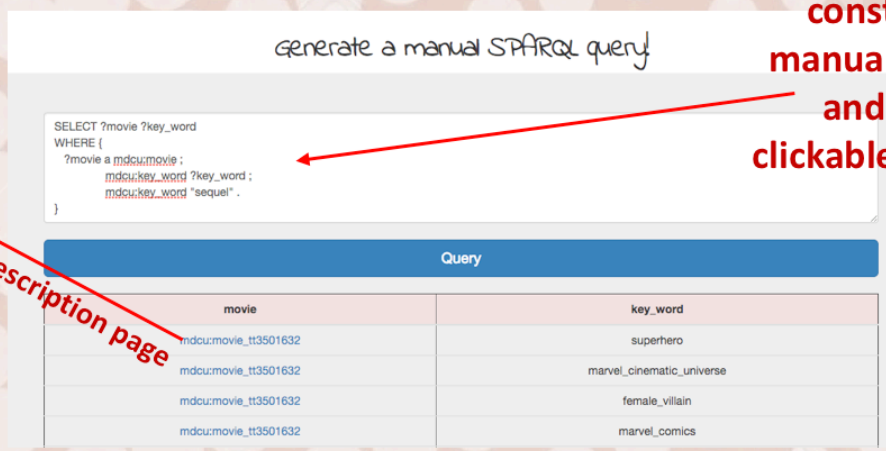
construct manual query and get clickable results





labels	values
mdcu:occupation_adventurer	1532
mdcu:occupation_criminal	1161





Validation

How to validate correctness of the KG in each 'curation' step?

generate error dump file with known patterns → cluster problems by type → manipulate using bash scripts (grep, cut, etc..)

Data Extraction

Data Cleaning

+

set of SPARQL queries → count and distinguish → triples by patterns

Data Cleaning

Triples curation

+

manual labeling of entities (validation set)

Entity Linking

+

online RDF validation tools

Triples logic

Ontology correctness

+

visualization tools

Ontology correctness

Evaluation

Fixes & deploy new system

Retrospective

pitch the project proposal to friends and colleagues

Present system

Feedback!
Satisfaction?
Interaction?

In the final version we presented users responded positively to all system components

Some users interacted with the system for more than 15 minutes

Use Cases

Deploy to external KGs (Linked Data)

contribute to the open world of linked data

Define axioms Run inference

utilize the information we have on relations between characters!

Predict! use ML!

good vs. bad, how is color related to that?
use comic-book scores and movie ratings!

Provider?

Explore the trends, the gaps, how can you fit in?

characters origins, popular attributes, ratings and scores

Consumer?

Navigate and walk through our graph! Explore further!

who is the inker of your favorite comic books issues?
what else did he ink?

Conclusions

Carefully design a focused crawler

don't 'blind' crawl

Working with existing ontologies can make it much easier

make use of work done by others!

Good ontology = Good KG

we believe that a good indicator for a well-constructed-KG is the complexity of composing a SPARQL query for any request

Blocking Schemes for Record Linkage = super useful

Can save a lot of time, even when it's not smart

Conclusions

KG > RDB

KGs shine when the goal is to capture complex relationships in vast webs of information.

In the domain we chose, and to the best of our knowledge, there is no active system which gives a user an easy way to navigate between the different personas behind the comic-book industry.

Our system makes that task look easy!