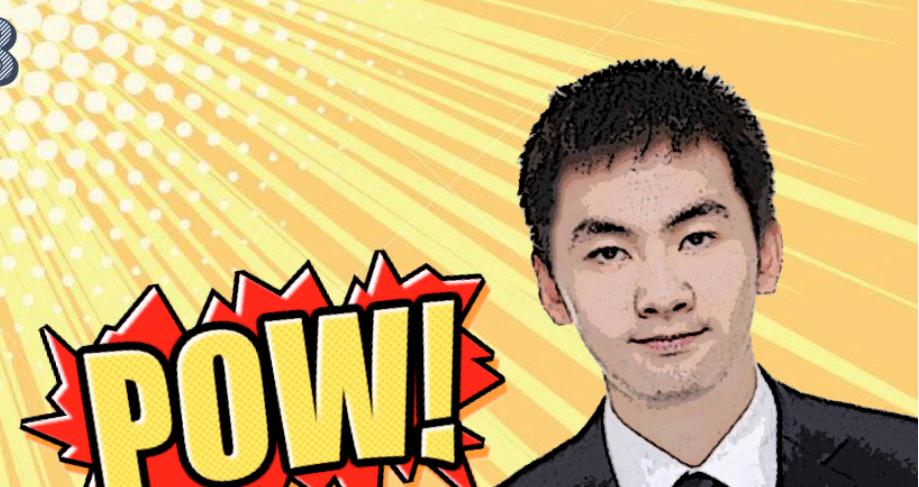


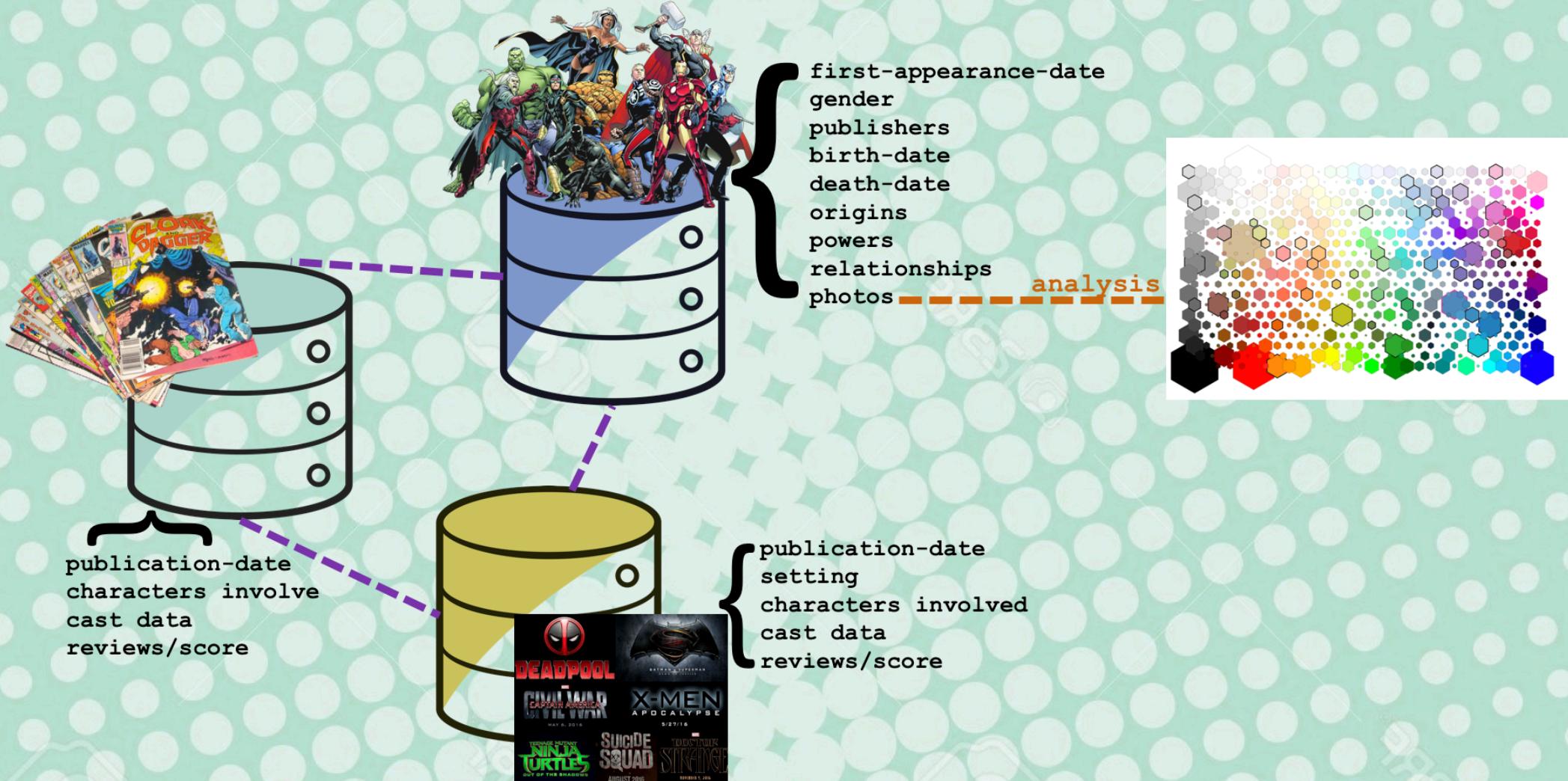
INF 558 Project Summary: Marvel and DC Comic Books Knowledge Graph

Basel Shbita, Ke Xu

11/30/18



Intro (Reminder)



Goals (Reminder)

Discover
New and interesting facts!

Search
More easily!



Expand

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

Analyze

Data!
Characters, movies & comic-book issues

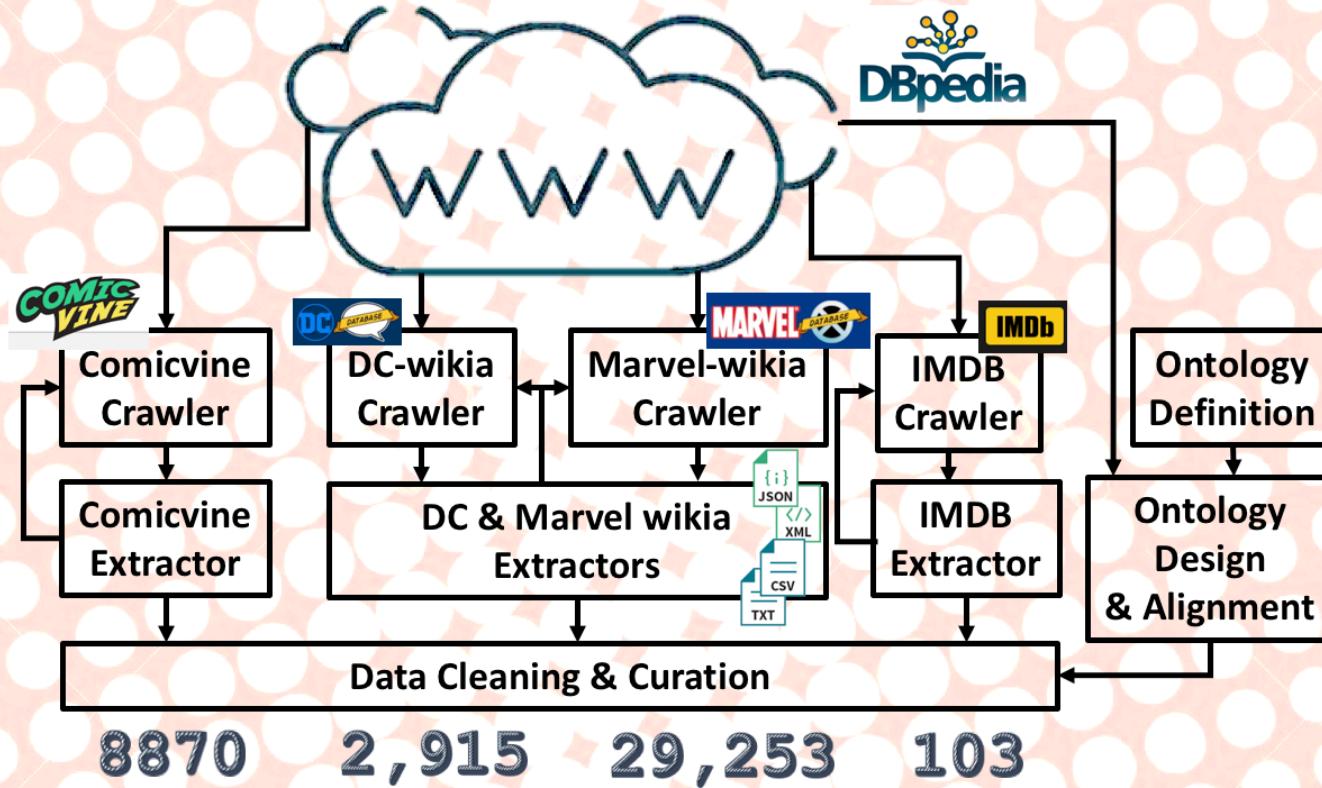
Explore

Further in their favorite domain

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

Earth-616 Characters

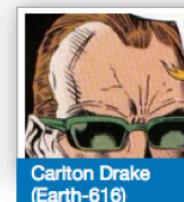
Category page

Characters from Earth-616.

0-9 • A • B • C • D • E • F • G • H • I • J • K • L • M • N • O • P • Q • R • S • T • U • V • W



Kenji Oyama
(Earth-616)



Carlton Drake
(Earth-616)



Kingmaker (Earth-616)



Ikari (Earth-616)



Cyttorak (Earth-616)



Jack O'Lantern
(Crime-Master's)
(Earth-616)



Gall (Earth-616)



Jonathan Murdock
(Earth-616)



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

Crawling Challenges

Problem Irrelevant Pages

Solution Pattern Matching

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

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

★★★★★ **Zombie Goodness..with a Twist**
68 Jungle Jim #1 - un by director421 on April 03, 2013
Highly detailed wor' artist Jeff Zornow Jim.This is the fi on, I am only s one-man kill!

★★★★★ **Introducing Mazing Man**
Mazing Man #1 by englentine on July 16, 2009
The book opens with Denton and his friend "Mazing Man walking through their neighborhood in New York City. Denton is an odd little man in that he looks like a dog. He's not a dog. He is a writer, and his friend is the self appointed protector of the neighborhood. A three foot harmless little man with a helmet and a cape. A few reviews ago, I called the Comet "The Best Kept Secret of the 90's" well I will make that announcement again for this failed series being the best kept secret of the 80's. ...



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

Extraction Challenges

Problem Structure Inconsistency

Rule-Based System

fr

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 [\[show\]](#)

System

Hist
fr
origin
ent
ic
re

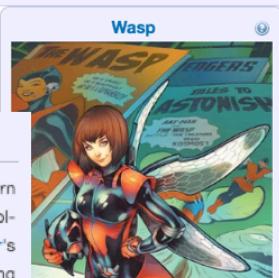
up in who was born into captivity and followed in her father's footsteps by mastering the Pym Particle, ultimately escaping to become the new [Wasp](#).

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



Origin

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

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

Data Cleaning Challenges

Problem
**Format
Inconsistency**

Solution
**Classify &
Transform**

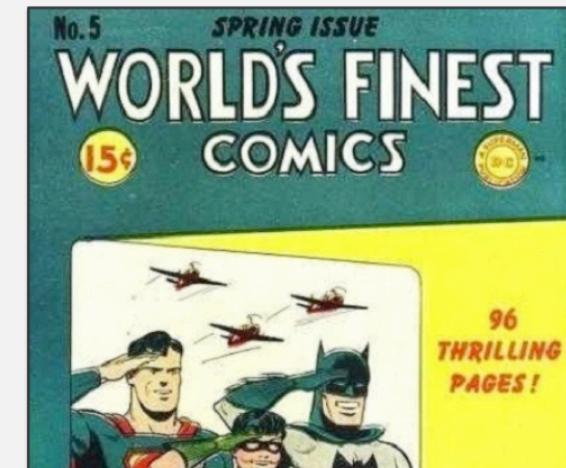
Classification determines the
Transformation type



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
Date	

Data Cleaning Challenges

Problem
**Multiple Values
Same Meaning**

Solution
Clustering

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

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

1,802 entries !!!

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

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

Ontology Challenges

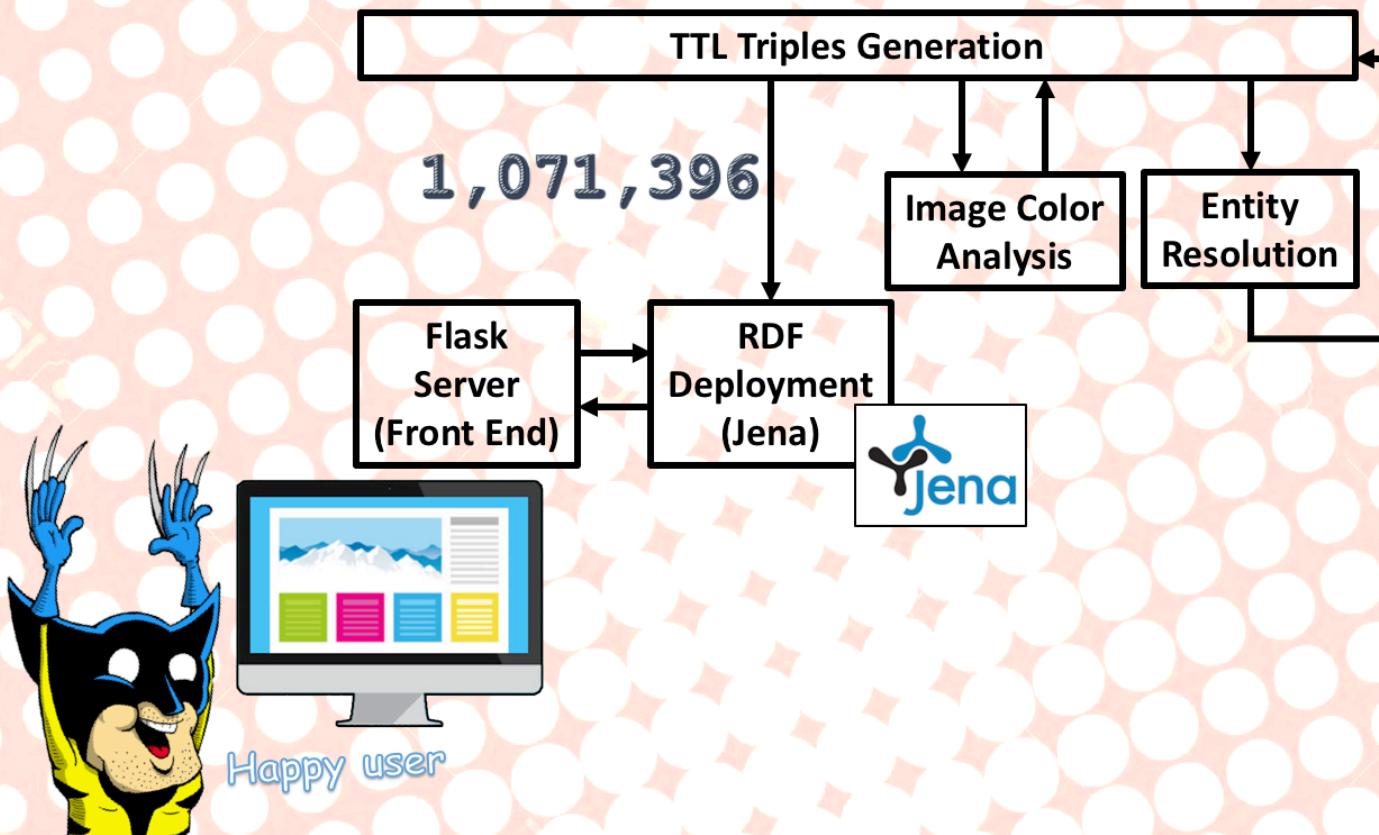
Problem
**Capture concepts
With minimal effort**

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

Movie.character -?> **Character**

Issue.character

Issue.location -?> **Character**.
 birth_place

place_of_operation

Issue.team -?> **Character.alliance**

Solution
ISI-RLTK

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

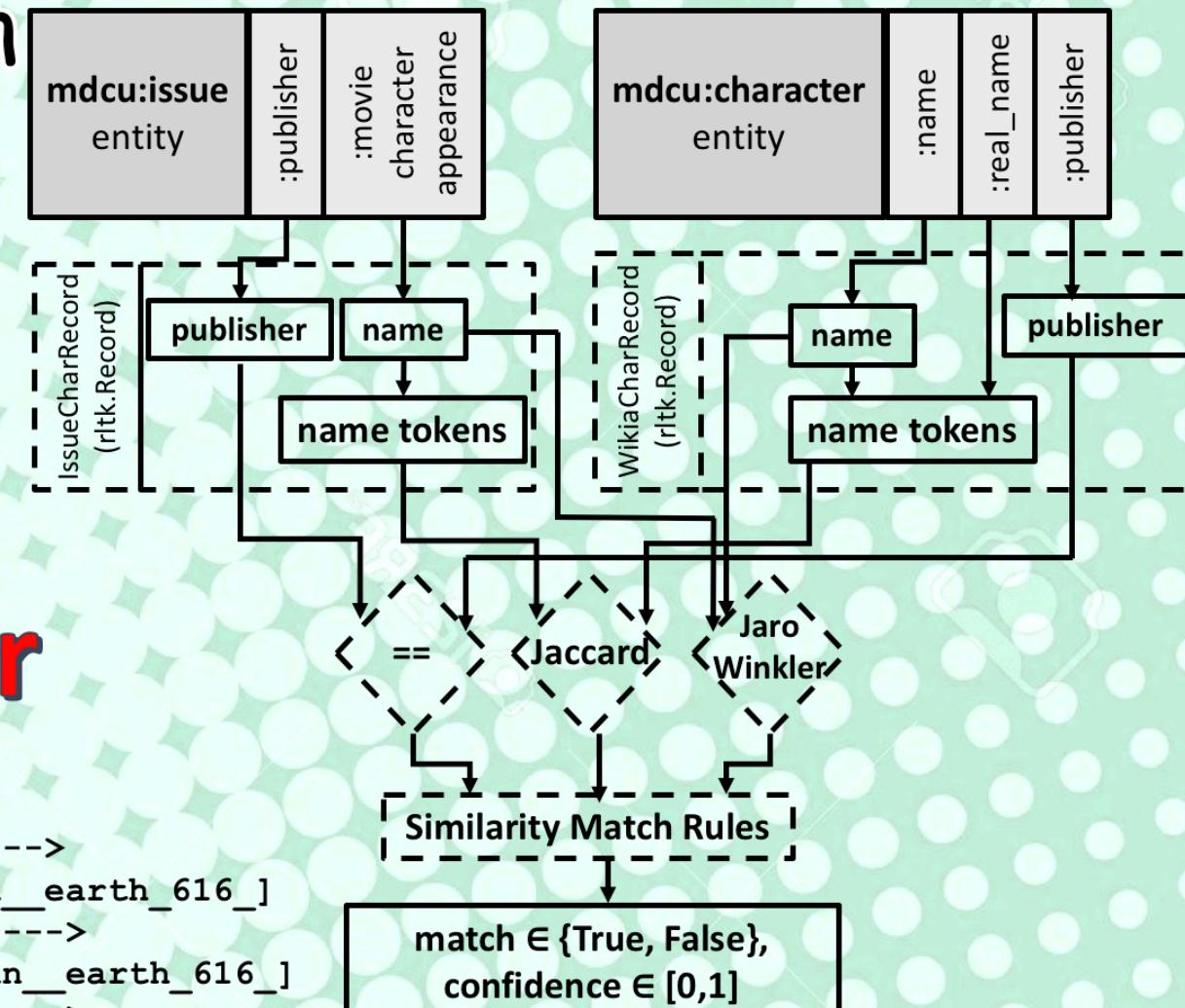
Entity Resolution

EXAMPLE

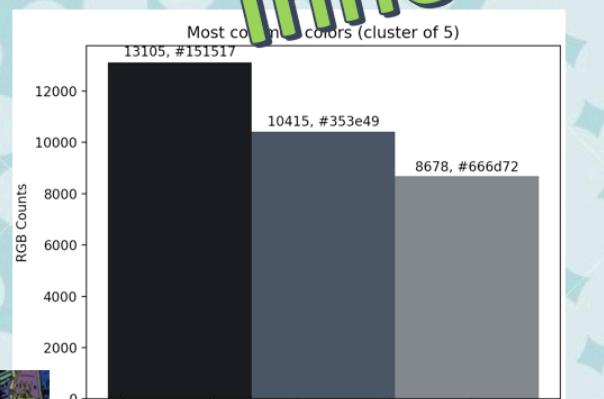
Issue.character

Character

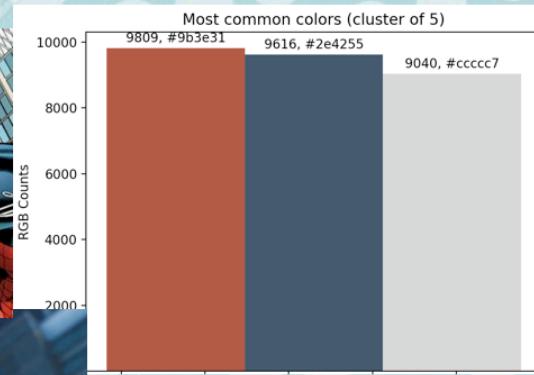
```
[mdcu:_morgan-stark_4005-19688] ---96.19%--->  
[mdcu:_wiki_morgan_stark_earth_616]  
[mdcu:_heather-glenn_4005-10365] ---99.17%--->  
[mdcu:_wiki_heather_glenn_earth_616]  
[mdcu:_richard-nixon_4005-13426] ---77.50%--->  
[mdcu:_wiki_richard_nixon_new_earth]
```



Color Analysis Innovation



['#151517', '#353e49', '#666d72']
[('black', (0, 0, 0)), 23520]



['#9b3e31', '#2e4255', '#cccccc']
[('red', (255, 0, 0)), 9809),
('blue', (0, 0, 255)), 9616]

For each character:
^{>30,000}

download image
(given img-url)

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)

resize

User Interface



Flask + + Bootstrap
+ Jena

Marvel and DC Comic Books KG

links to a description page

links to a description page

perform text search within results

sort by s,p,o

pagination

subject	predicate	object
mdcu_wiki_27electron_earth_12772...	mdcu:has_occupation	mdcu:occupation_adventurer
mdcu_wiki_a_n_2010_earth_tm157...	mdcu:has_occupation	mdcu:occupation_adventurer
mdcu_wiki_sarkus_earth_9997...	mdcu:has_occupation	mdcu:occupation_adventurer
mdcu_wiki_abraham_brown_earth_616...	mdcu:has_occupation	mdcu:occupation_adventurer
mdcu_wiki_ace_g_r_a_m_p_a_earth_616...	mdcu:has_occupation	mdcu:occupation_adventurer
mdcu_wiki_alpha_earth_616...	mdcu:has_occupation	mdcu:occupation_adventurer
mdcu_wiki_ami_yoshida_earth_12131...	mdcu:has_occupation	mdcu:occupation_adventurer
mdcu_wiki_iges27ha_earth_616...	mdcu:has_occupation	mdcu:occupation_adventurer
mdcu_wiki_aldrit_odinadottir_earth_tm517...	mdcu:has_occupation	mdcu:occupation_adventurer
mdcu_wiki_alison_blaire_earth_616...	mdcu:has_occupation	mdcu:occupation_adventurer

Show 10+ entries

search:

Results

Showing 1 to 10 of 203 entries

Prev 1 2 3 4 5 ... 21 Next

main page

Top Results of http://inf558.org/comics#has_occupation for mdcucharacter

Choose a class: [mdcu:character](#) [http://inf558.org/comics#has_occupation](#)

predicates are generated automatically in BE

responsive pie and bar charts generated automatically

construct manual query and get clickable results

dropdown menu for class selection

generate a manual SPARQL query

```
SELECT ?movie ?key_word
WHERE {
?movie a mdcu:movie ;
      mdcu:key_word ?key_word ;
      mdcu:key_word "sequel" .
}
```

Query

movie	key_word
mdcu:movie_tt3501632	superhero
mdcu:movie_tt3501632	marvel_cinematic_universe
mdcu:movie_tt3501632	female_villain
mdcu:movie_tt3501632	marvel_comics

http://inf558.org/comics#role_char_creator
http://inf558.org/comics#provenance_url
http://inf558.org/comics#has_color
http://inf558.org/comics#has_identity
http://inf558.org/comics#hair
http://xmlns.com/foaf/0.1/gender
http://inf558.org/comics#first_appearance
http://inf558.org/comics#has_citizenship
http://inf558.org/comics#operation_base
http://inf558.org/comics#marital_status
http://dbpedia.org/property/alliances
http://dbpedia.org/property/aliases
http://inf558.org/comics#issue_appearance
http://inf558.org/comics#has_alignment
http://inf558.org/comics#eyes
http://schema.org/birthPlace
http://inf558.org/comics#has_ability
http://dbpedia.org/property/powers
✓ http://inf558.org/comics#has_occupation

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

pitch the project proposal to friends and colleagues

*Fixes & deploy new system
Retrospective*

Feedback!
Satisfaction?
Interaction?

Present system



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

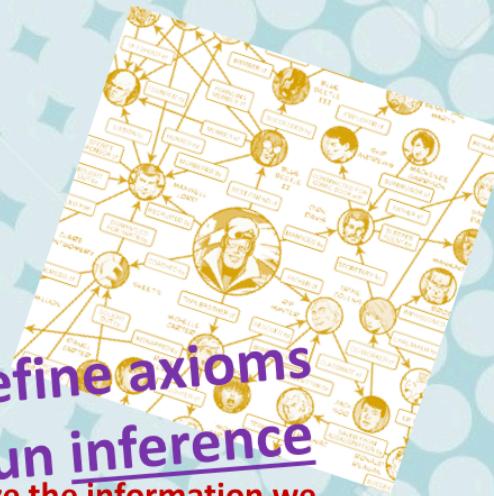
The diagram features a large purple title at the top left: "Deploy to external KGs (Linked Data)". Below it, a red subtitle reads: "contribute to the open world of linked data". To the right, a network graph shows various entities as nodes connected by lines representing relationships. Nodes include "names", "Eurostat", "G", "World Fact-book", "DBpedia", "Magna Graiae", "se", "US Census Data", "3C WorldNet", and "UMBEL". A legend at the bottom right defines symbols: a blue circle for "linked data", a green circle for "linked data", and a red circle for "linked data".



Provider?
Explore the trends,
the gaps, how can
you fit in?
characters origins, popular
attributes, ratings and scores



Define axioms
Run inference
utilize the information we
have on relations between
characters!



Consumer?
Navigate and walk
through our graph!
Explore further!

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



Predict! use ML!

use ML!

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

Conclusions

Carefully design a focused crawler
don't 'blind' crawl

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



make use of work done by others!

Working with existing ontologies can make it much easier

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!

