

Science Stack Exchange

Aanchal Mahajan, Kaustuv Deolal, Mahima Gupta, Sanchari Banerjee (Team-13)



Table of contents

- Agenda
- Inspiration
- Data Domains
- Semantic Data Model
- Graph Diagram
- Linked Data Generation
- Our Plan with Linked Data
- Work Flow
- Teamwork and Collaboration

Agenda



The project aims at creating a knowledge tool for topics in pure sciences viz. physics, chemistry, mathematics and biology. Every upcoming researcher backtracks to find solution to common problems. Most of the times he is directed to redundant sources and ends up wasting his precious time. This resource would aim at delivering precise results to user queries in the aforementioned domains.

Inspiration

All Queries

293 views

How many upvotes do I have for each tag?

sep 29 11 sam.saffron

212 views

How Unsung am I?

oct 4 11 Eric

173 views

My Comment Score distribution

dec 22 11 sam.saffron

142 views

What is my accepted answer percentage rate

oct 3 11 sam.saffron

118 views

Find interesting unanswered questions

oct 3 11 sam.saffron

104 views

StackOverflow Rank and Percentile

nov 4 12 Cade Roux

81 views

My Money for Jam

oct 4 11 krock

featured recent

Choose a Site

Data updated 13 hours ago



Stack Overflow
Q&A for programmers

15m questions 23m answers 62m comments 51k tags visit site →



Mathematics
Q&A for people studying math at any level and professionals in related fields

867k questions 1.2m answers 3.8m comments 1.7k tags visit site →



Super User
Q&A for computer enthusiasts and power users

363k questions 539k answers 1.3m comments 5.2k tags visit site →



Ask Ubuntu
Q&A for Ubuntu users and developers

278k questions 360k answers 949k comments 3k tags visit site →



Server Fault
Q&A for system administrators and IT professionals

253k questions 421k answers 833k comments 3.6k tags visit site →



Stack Overflow на русском
Q&A for programmers

179k questions 216k answers 745k comments 3.9k tags visit site →



TeX - LaTeX
Q&A for users of TeX, LaTeX, ConTeXt, and related typesetting system

146k questions 191k answers 765k comments 1.6k tags visit site →

How is our
application **different?**

- We concentrate on queries related to science and hacking only
- We have datasets from four different domains not just stack exchanges
- We put emphasis on obtaining knowledge not just redundant statistics
- Our application does not concentrate on one User
- The data sources used cover both modern day and old school science.

Data Domains

Hackernews

Hacker News is a social news website focusing on computer science and start-ups. The intention was to recreate a community similar to the early days of Reddit.

 **Hacker News** new | comments | show | ask | jobs | submit

1. ▲ [The beginning of the end for copper wire](#) (potsandpansbyccg.com)
71 points by rmason 1 hour ago | hide | 53 comments
2. ▲ [Lisp in fewer than 200 lines of C](#) (carl.github.io)
150 points by jfo 4 hours ago | hide | 40 comments
3. ▲ [Facebook Is the Junk Food of Socializing \(2015\)](#) (nautil.us)
75 points by dnetesn 4 hours ago | hide | 4 comments
4. ▲ [Rust: Enable WebAssembly backend by default](#) (hellorust.com)
327 points by a_humeau 8 hours ago | hide | 125 comments
5. ▲ [HAProxy 1.8.0](#) (mail-archive.com)
46 points by rjgray 1 hour ago | hide | 5 comments
6. ▲ [Can anyone make money on the moon?](#) (nytimes.com)
18 points by jaredwiener 1 hour ago | hide | 5 comments
7. ▲ [Super Tiny Website Logos In SVG](#) (shkspr.mobi)
285 points by edent 9 hours ago | hide | 73 comments
8. ▲ [Some chip makers have hidden latency and jitter issues from common tests](#) (badmoderns.com)
128 points by basetd2 5 hours ago | hide | 20 comments
9. ▲ [A Deep Dive into NEC's Aurora Vector Engine](#) (nextplatform.com)
40 points by rbanffy 3 hours ago | hide | 3 comments
10. ▲ [Microsoft doubles down on Kubernetes for Azure](#) (dxc.technology)
25 points by CrankyBear 2 hours ago | hide | 17 comments
11. ▲ [A friendly first-principles intro to backprop in python](#) (sushant-choudhary.github.io)
39 points by sushantic 3 hours ago | hide | 4 comments
12. ▲ [Start Tracking Satellites with This Low-Cost Azimuth-Elevation Positioner](#) (hackaday.com)
24 points by LarryMandhane 3 hours ago | hide | 2 comments
13. ▲ [Revisiting the Mutilated Chessboard](#) (solipsys.co.uk)
29 points by ColinWright 5 hours ago | hide | 6 comments
14. ▲ [Firefox will soon flag sites that have been hacked](#) (engadget.com)
83 points by joeyespo 4 hours ago | hide | 26 comments
15. ▲ [Bitcoin Blows Past \\$9,000](#) (gizmodo.com)
68 points by ourmandave 3 hours ago | hide | 45 comments
16. ▲ [Physicists make most precise measurement ever of the proton's magnetic moment](#) (phys.org)
69 points by dnetesn 9 hours ago | hide | 43 comments
17. ▲ [A CMS with no server and 18 lines of configuration](#) (netlify.com)
113 points by owennm 12 hours ago | hide | 60 comments
18. ▲ [Photovoltaic growth: reality versus projections of the IEA – the 2017 update](#) (steinbuch.wordpress.com)
10 points by mervinmills 3 hours ago | hide | 1 comment

Science Stack Exchange

We are integrating datasets from four stack exchanges viz. Maths, Physics , Chemistry and Biology.

This screenshot shows the homepage of the Human Biology Stack Exchange. It features a sidebar with tags like 'human-biology', 'genetics', 'evolution', etc., and a main area with three recent questions:

- 0 votes, 0 answers, 8 views: Could I be pregnant? (asked 19 mins ago by user1)
- 2 votes, 0 answers, 4 views: Are dead ants returned to colony? (asked 1 hour ago by antbot)
- 2 votes, 1 answer, 23 views: Identify this plant (answered 1 hour ago by Kurt) (species-identification, botany)

(sequences-and-series) (combinatorics) (general-topology) (matrices) more tags

This screenshot shows a question from the Mathematics Stack Exchange. The question is: "Lower bound for the size of the maximum matching in a particular bipartite graph" (graph-theory, matching-theory). It has 0 votes, 0 answers, and 2 views. It was asked 1 min ago by Pedro Alves.

This screenshot shows a question from the Mathematics Stack Exchange. The question is: "Proving taylor polynomial of $\cos(x^2)$ of degree $4n$ " (polynomials, taylor-expansion). It has 0 votes, 0 answers, and 19 views. It was modified 1 min ago by user377299.

This screenshot shows a question from the Mathematics Stack Exchange. The question is: "Question using Dirichlet Approximation Theorem" (number-theory). It has 0 votes, 0 answers, and 8 views. It was modified 1 min ago by dmsj.

This screenshot shows the homepage of the Chemistry Stack Exchange. It features a sidebar with tags like 'organic-chemistry', 'homework', etc., and a main area with two recent questions:

- 0 votes, 1 answer, 17 views: What can we deduce about a (complex, organic) molecule's structure from the order of elements in its formula? (organic-chemistry, nomenclature) (answered 23 mins ago by TheChemist)
- 1 vote, 1 answer, 18 views: Do 1s & 2s orbitals overlap of different atoms while forming sigma bond? I (orbitals, molecular-structure, valence-bond-theory) (modified 50 mins ago by Oscar Lanzi)

This screenshot shows the homepage of the Physics Stack Exchange. It features a sidebar with tags like 'quantum-mechanics', 'homework-and-exercises', etc., and a main area with two recent questions:

- 0 votes, 0 answers, 22 views: Expectation value of Square Momentum (quantum-mechanics, operators, momentum, wavefunction, hilbert-space) (modified 5 mins ago by NegativeTensor)
- 0 votes, 1 answer, 15 views: What has more force/power? (forces) (answered 5 mins ago by Floris)

This screenshot shows a question from the Biology Stack Exchange. The question is: "Simulate and interpret a simulated SUSY event from Atlas" (particle-physics, quantum-field-theory, thermodynamics, general-relativity, special-relativity, classical-mechanic, forces). It has 0 votes, 0 answers, and 4 views. It was modified 7 mins ago by Floris.

Book-Crossing Dataset

This is a 4-week crawl from [Book-Crossing](#). It contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.



[Institut für Informatik, Universität Freiburg](#)

Book-Crossing Dataset ... mined by [Cai-Nicolas Ziegler, DBIS Freiburg](#)

Collected by Cai-Nicolas Ziegler in a 4-week crawl (August / September 2004) from the [Book-Crossing](#) community with kind permission from Ron Hornbaker, CTO of [Humankind Systems](#). Contains 278,858 users (anonymized but with demographic information) providing 1,149,780 ratings (explicit / implicit) about 271,379 books.

[!] Freely available for research use when acknowledged with the following reference (further details on the dataset are given in this publication):

- [Improving Recommendation Lists Through Topic Diversification](#),
Cai-Nicolas Ziegler, Sean M. McNee, Joseph A. Konstan, Georg Lausen; *Proceedings of the 14th International World Wide Web Conference (WWW '05)*, May 10-14, 2005, Chiba, Japan. To appear.

Download: [[PDF Pre-Print](#)]

As a courtesy, if you use the data, I would appreciate knowing your name, what research group you are in, and the publications that may result.

Format

The Book-Crossing dataset comprises 3 tables.

- BX-Users
Contains the users. Note that user IDs ('User-ID') have been anonymized and map to integers. Demographic data is provided ('Location', 'Age') if available. Otherwise, these fields contain NULL-values.
- BX-Books
Books are identified by their respective ISBN. Invalid ISBNs have already been removed from the dataset. Moreover, some content-based information is given ('Book-Title', 'Book-Author', 'Year-Of-Publication', 'Publisher'), obtained from Amazon Web Services. Note that in case of several authors, only the first is provided. URLs linking to cover images are also given, appearing in three different flavours ('Image-URL-S', 'Image-URL-M', 'Image-URL-L'), i.e., small, medium, large. These URLs point to the Amazon web site.
- BX-Book-Ratings
Contains the book rating information. Ratings ('Book-Rating') are either explicit, expressed on a scale from 1-10 (higher values denoting higher appreciation), or implicit, expressed by 0.

MathML- Journals

The dataset is an archive of articles from journals which are organized into web pages. The dataset is in XML format.

```
- <ref-list>
+ <ref id="B1" content-type="article" article_id="2231.0B1">
+ <ref id="B2" content-type="misc" article_id="2231.0B2">
+ <ref id="B3" content-type="article" article_id="2231.0B3">
+ <ref id="B4" content-type="article" article_id="2231.0B4">
+ <ref id="B5" content-type="article" article_id="2231.0B5">
+ <ref id="B6" content-type="article" article_id="2231.0B6">
+ <ref id="B7" content-type="article" article_id="2231.0B7">
+ <ref id="B8" content-type="article" article_id="2231.0B8">
+ <ref id="B9" content-type="article" article_id="2231.0B9">
+ <ref id="B10" content-type="article" article_id="2231.0B10">
+ <ref id="B11" content-type="article" article_id="2231.0B11">
+ <ref id="B12" content-type="article" article_id="2231.0B12">
+ <ref id="B13" content-type="article" article_id="2231.0B13">
+ <ref id="B14" content-type="article" article_id="2231.0B14">
+ <ref id="B15" content-type="article" article_id="2231.0B15">
+ <ref id="B16" content-type="article" article_id="2231.0B16">
+ <ref id="B17" content-type="article" article_id="2231.0B17">
+ <ref id="B18" content-type="article" article_id="2231.0B18">
- <ref id="B19" content-type="article" article_id="2231.0B19">
```

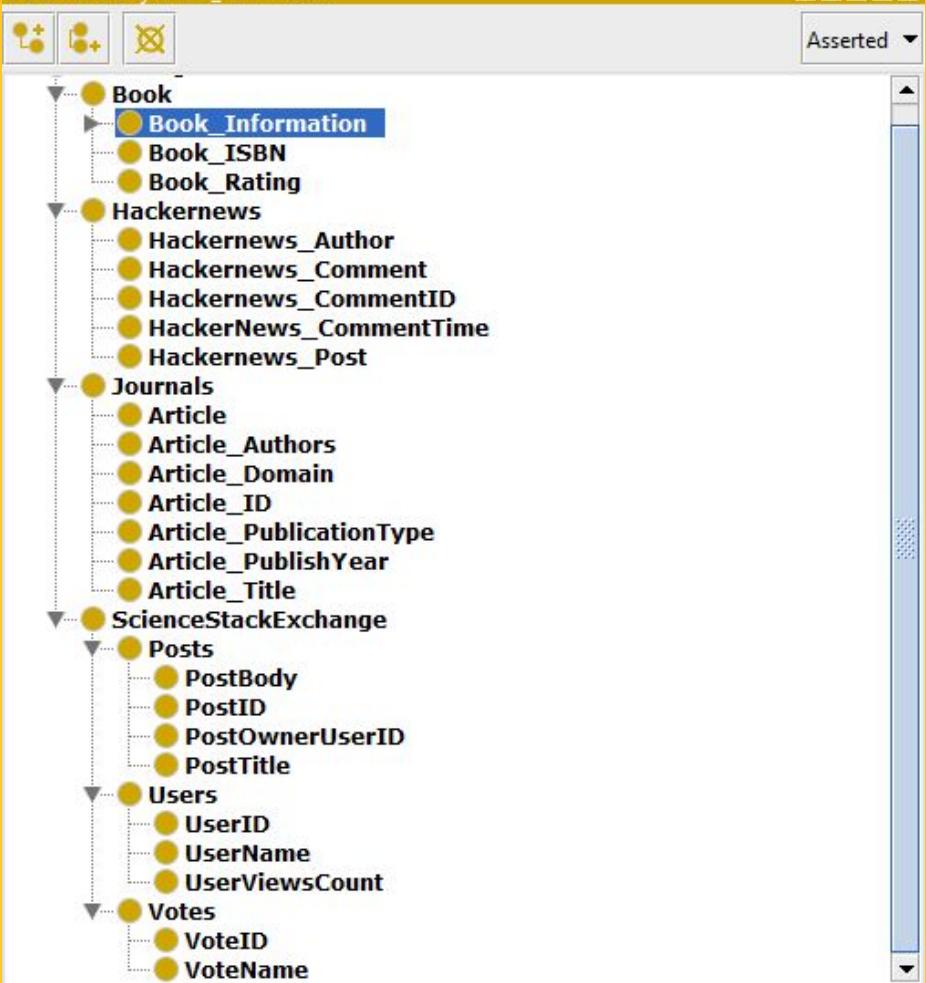
Every XML file has a ref-list under which various articles with unique article_id are mentioned.

```
- <ref id="B1" content-type="article" article_id="2231.0B1">
  - <nlm-citation publication-type="journal">
    - <person-group person-group-type="author">
      - <name>
        <surname>Steinmann</surname>
        <given-names>P.</given-names>
      </name>
      - <name>
        <surname>Keiser</surname>
        <given-names>J.</given-names>
      </name>
      - <name>
        <surname>Bos</surname>
        <given-names>R.</given-names>
      </name>
      - <name>
        <surname>Tanner</surname>
        <given-names>M.</given-names>
      </name>
      - <name>
        <surname>Utzinger</surname>
        <given-names>J.</given-names>
      </name>
    </person-group>
    <article-title>Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk</article-title>
    - <source>
      <italic>Lancet Infectious Diseases</italic>
    </source>
    <year>2006</year>
```

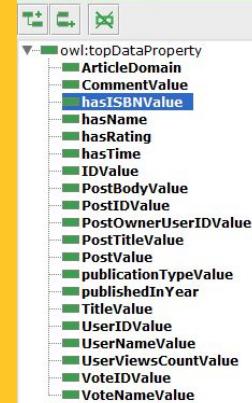
Under every ref-id tag, relevant information is present under other tags like publication-type, author name (divided into given-names and surname), article-title, source (journal name) and year of publication is mentioned.

Semantic Data Model

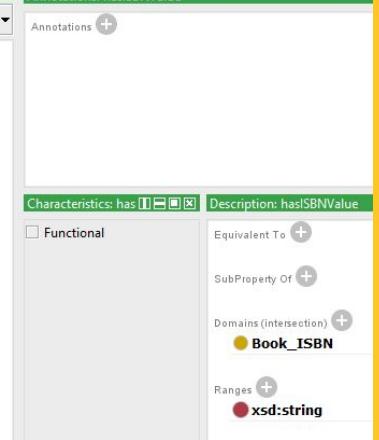
Class hierarchy: Book_Information



Data property hierarchy: hasISBNValue



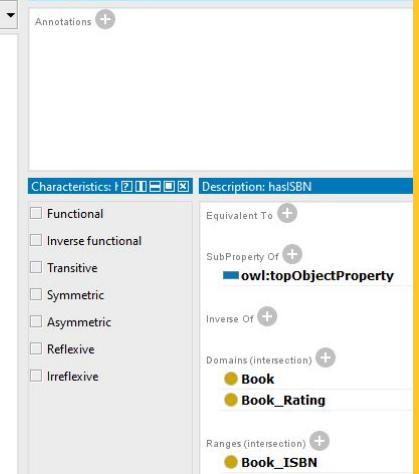
Annotations: hasISBNValue



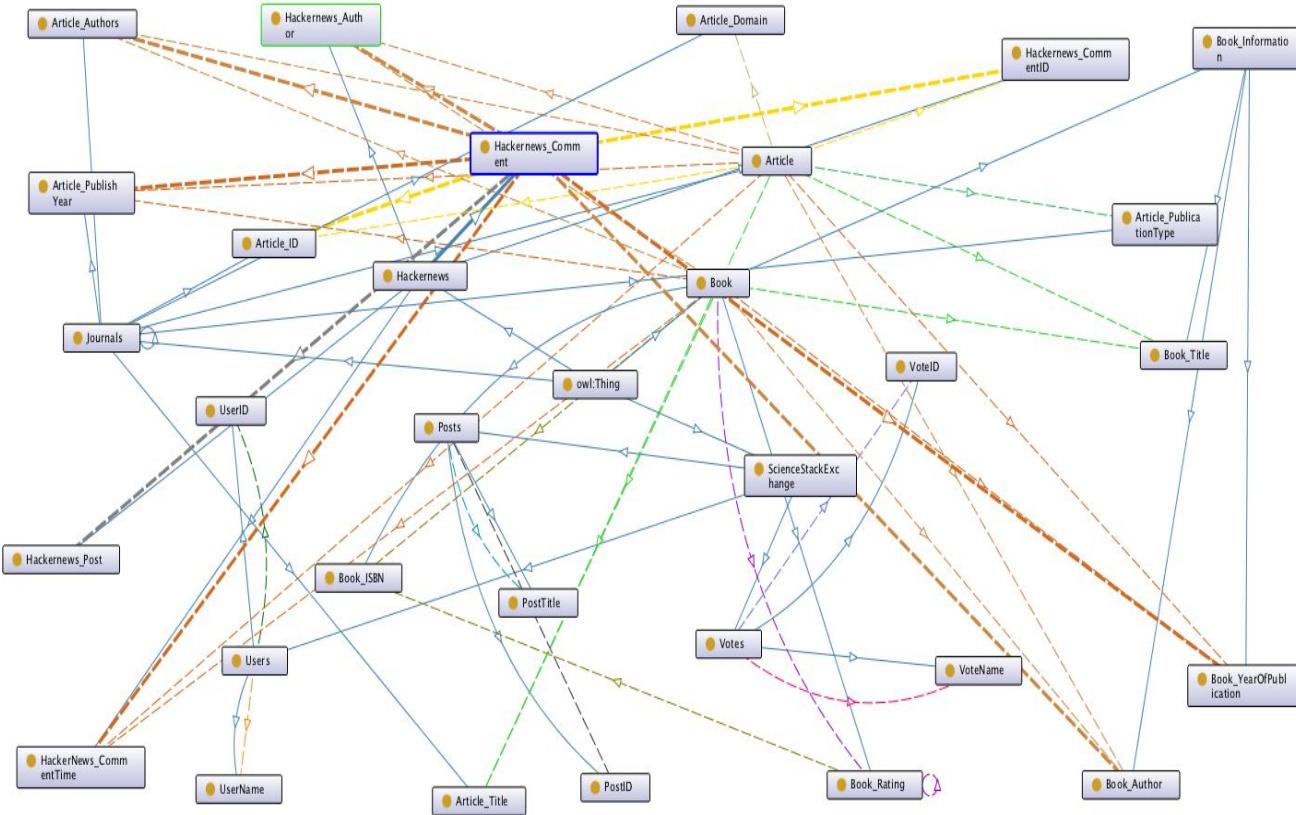
Object property hierarchy: hasISBN



Annotations: hasISBN



Graph Diagram



-  belongsTo (Domain>Range)
 -  from (Domain>Range)
 -  has individual
 -  has subclass
 -  hasID (Domain>Range)
 -  hasISBN (Domain>Range)
 -  hasPostID (Domain>Range)
 -  hasPostTitle (Domain>Range)
 -  hasRating (Domain>Range)
 -  hasTitle (Domain>Range)
 -  hasUserID (Domain>Range)
 -  hasUserName (Domain>Range)
 -  hasVotID (Domain>Range)
 -  hasVoteName (Domain>Range)
 -  originatedFrom (Domain>Range)
 -  publishedIn (Domain>Range)
 -  writtenBy (Domain>Range)

Process of Linked Data Generation

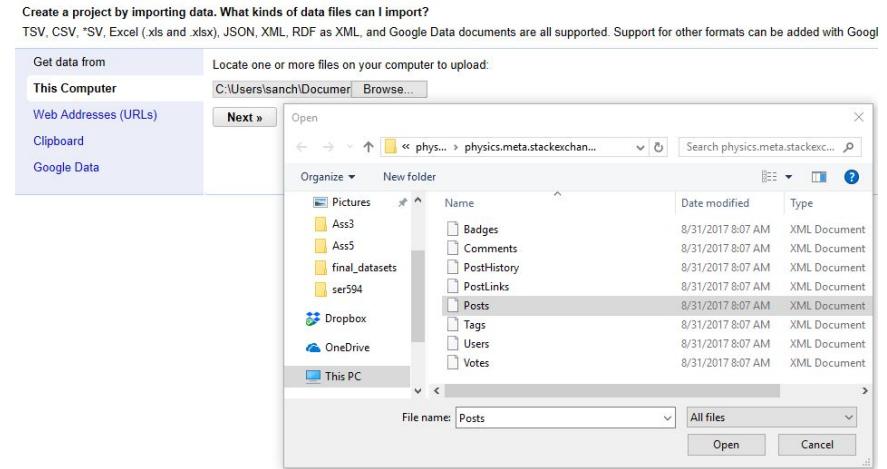
Hackernews

- Converted JSON Dataset to CSV Using pandas in python
- Created rdf Triples using Karma.
- Uploaded the csv and owl file on the karma server and mapped the properties with classes

```
import pandas as pd
df=pd.read_json("hc_news_2017-05-01.json")
print(df)
df.to_csv('hc_news_2017_05_01.csv')
```

Science Stack Exchange

- Uploaded the XML datasets on Google Refine
- Added the respective OWL(sciencestackexchange.owl) file
- Mapped the XML dataset field names/attributes (postID, postTitle, UserID etc.) to the object properties in OWL file
- Generated RDF triples by exporting the output as RDF turtle



Book-Crossing

- Uploaded the crawled CSV datasets on Google Refine from Book-Crossing.
- Added the respective OWL(sciencestackexchange.owl) file with a prefix:"sse"
- Mapped the CSV dataset field names/attributes (ISBN, BookTitle, UserID etc.) with the object properties in OWL file
- Generated RDF triples by exporting the output as RDF turtle.

Google refine A power tool for working with messy data.

Create Project « Start Over Configure Parsing Options Project name BX Books.csv Create Project »

Open Project Import Project

	ISBN	Book-Title	Book-Author	Year-Of-Publication	Publisher	Image-URL-S
1.	195153448	Classical Mythology	Mark P. O. Morford	2002	Oxford University Press	http://images.amazon.com/images/P/0195153448.01.THUMBZZ
2.	2005018	Clara Callan	Richard Bruce Wright	2001	HarperFlamingo Canada	http://images.amazon.com/images/P/0002005018.01.THUMBZZ
3.	60973129	Decision in Normandy	Carlo D'Este	1991	HarperPerennial	http://images.amazon.com/images/P/0060973129.01.THUMBZZ
4.	374157065	Flu: The Story of the Great Influenza Pandemic of 1918 and the Search for the Virus That Caused It	Gina Bari Kolata	1999	Farrar Straus Giroux	http://images.amazon.com/images/P/0374157065.01.THUMBZZ
5.	393045218	The Mummies of Urumchi	E. J. W. Barber	1999	W. W. Norton & Company	http://images.amazon.com/images/P/0393045218.01.THUMBZZ
6.	399135782	The Kitchen God's Wife	Amy Tan	1991	Putnam Pub Group	http://images.amazon.com/images/P/0399135782.01.THUMBZZ
7.	425176428	What If?: The World's	Robert Cowley	2000	Berkley Publishing Group	http://images.amazon.com/images/P/0425176428.01.THUMBZZ

BX-Books-csv.ttl

```
1 @prefix sse: <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#> .
2 @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
3 @prefix owl: <http://www.w3.org/2002/07/owl#> .
4 @prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
5 @prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
6 @prefix foaf: <http://xmlns.com/foaf/0.1/> .

7
8
9 <http://127.0.0.1:3333/0> a <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#Book> ;
10 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#hasISBN> "195153448" ;
11 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#hasTitle> "Classical Mythology" ;
12 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#writtenBy> "Mark P. O. Morford" ;
13 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#publishedIn> "2002" ;
14 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#publishedby> "Oxford University Press" .
15
16 <http://127.0.0.1:3333/1> a <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#Book> ;
17 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#hasISBN> "2005018" ;
18 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#hasTitle> "Clara Callan" ;
19 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#writtenBy> "Richard Bruce Wright" ;
20 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#publishedIn> "2001" ;
21 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#publishedby> "HarperFlamingo Canada" .
22
23 <http://127.0.0.1:3333/2> a <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#Book> ;
24 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#hasISBN> "60973129" ;
25 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#hasTitle> "Decision in Normandy" ;
26 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#writtenBy> "Carlo D'Este" ;
27 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#publishedIn> "1991" ;
28 <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untilled-ontology-6#publishedby> "HarperPerennial" .
```

MathML- Journals

- Before I converted the dataset into RDF triples, I cleaned the data by removing tags which were not required.
- Unique article_id was also inserted which will be used a primary key during querying.
- Cleaning and Modification of data was performed by XML DOM Parser

```
- <front>
  + <journal-meta>
  + <article-meta>
</front>
- <body>
  + <sec id="sec1" sec-type="section">
  + <sec id="sec2" sec-type="section">
  - <sec id="sec3" sec-type="section">
    <title>3. Stability</title>
    + <p>
    + <statement id="lem2">
    + <statement id="proof3">
    + <statement id="ex3">
    + <p>
    + <statement id="thm2">
    + <statement id="proof4">
    + <p>
    + <p>
  </sec>
</body>
- <back>
  - <ref-list>
    + <ref id="B22" content-type="book">
    + <ref id="B6" content-type="article">
    + <ref id="B2" content-type="article">
    + <ref id="B3" content-type="article">
    - <ref id="B4" content-type="article">
      . . .
```

- Created a zip file of about 2000 files
- Uploaded the zip on Google Refine
- Added the respective OWL(sciencestackexchange.owl) file
- Mapped the XML dataset field names/attributes (articleID, authorName, articleTitle, yearOfPublication etc) with the object properties in OWL file
- Generated RDF triples by exporting the output as RDF turtle
- Repeated these steps as dataset contains over 100,000 files.

Last modified	Name
a week ago	t18 197491 2012 07 11 xml
a week ago	t18 197491 2012 07 11 xml
a week ago	t17 1315497 2017 01 12 xml
a week ago	t16 160174 2012 09 05 xml
a week ago	t15 1286315 2016 06 08 xml
a week ago	t13 104535 2014 06 19 xml
a week ago	t9 5026504 2016 06 09 xml
a week ago	t8 143471 2011 03 14 xml
a week ago	t4 101074 2012 05 06 xml
a week ago	journals 101542 2012 07 12 xml
a week ago	t 105875 2014 06 18 xml

Available Prefixes:

sse rdf owl xsd rdfs foaf + add prefix ⚙ manage prefixes

(row index) URI

×sse:Article

add rdf:type

×	→sse:hasID→	☒ ref-list - ref - article_id cell
×	→sse:from→	☒ ref-list - ref - nlm-citation - publication-type cell
×	→sse:writtenBy→	☒ ref-list - ref - nlm-citation - person-group - name - surname cell
×	→sse:writtenBy→	☒ ref-list - ref - nlm-citation - person-group - name - given-names cell
×	→sse:writtenBy→	☒ ref-list - ref - nlm-citation - person-group - name - suffix cell
×	→sse:hasTitle→	☒ ref-list - ref - nlm-citation - article-title cell
×	→sse:publishedIn→	☒ ref-list - ref - nlm-citation - year cell
×	→sse:domain→	☒ ref-list - ref - nlm-citation - source - italic cell

Mapping of dataset field to corresponding object property

```
<http://127.0.0.1:3333/0> a <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#Article> ;  
  <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#hasID> "1.0B1" ;  
  <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#from> "journal" ;  
  <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#writtenBy> "Dal-Pizzol" , "F." ;  
  <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#hasTitle> "Alternative activated macrophage: a new key for systemic inflammatory  
  <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#publishedIn> "2004" ;  
  rdfs:domain "Critical Care Medicine" .
```

RDF Triple in TTL format

Our Plan With Integrated Data

Queries

1. What is the atomic number of Manganese?
2. What is Einstein's law of relativity?
3. Who is the first American to walk on moon?
4. What is a factorial?
5. Research on various parameters of radio channels
6. Experimental results of resonance under different scenarios
7. Research on genetic systems
8. Research on reprogramming of chemical compounds

Queries specific to
Datasets

The Bigger Picture

For a novice, our application will provide comprehensive results. For example one wants to learn about the naming of orbits in atoms, our application will provide the relevant discussions from blogs, books that contain this information and research papers explaining the same and those articles that used this information to get other results.

WorkFlow

- The project has two parts:
- The majority of the frontend deals with users running SPARQL queries on integrated datasets.
- The queries run at the backend. Users don't need to write them per se
- The second part of the project concentrates on creating the APIs for the integrated dataset.
- Free API would be provided for customers with an authentication keys. These API would be relevant to the queries specific to the researchers and won't be displayed on the main web app.

Teamwork and Collaboration

- The work on this project was done according to Agile Methodology of Software development
- The team member met on a weekly basis discussing their ideas and the current development
- Since each of us has a different dataset from four varied domains, operations like data cleaning and generating RDF triples was individual work
- Efforts like designing the workflow and making the Ontology was a collaborative effort of all team members.
- Lastly the work on webapp was divided equally into the frontend, backend and data querying processes.

Thank You

