Science Stack Exchange

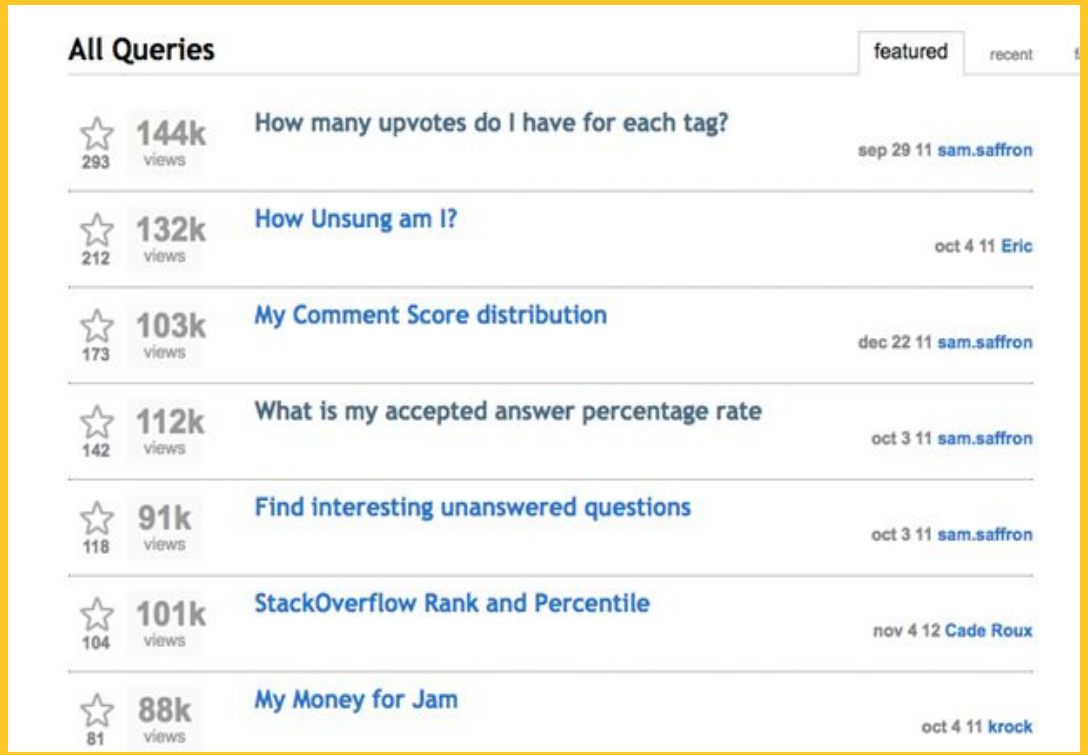
Aanchal Mahajan, Kaustuv Deolal, Mahima Gupta, Sanchari Banerjee (Team-13)

Table of contents

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



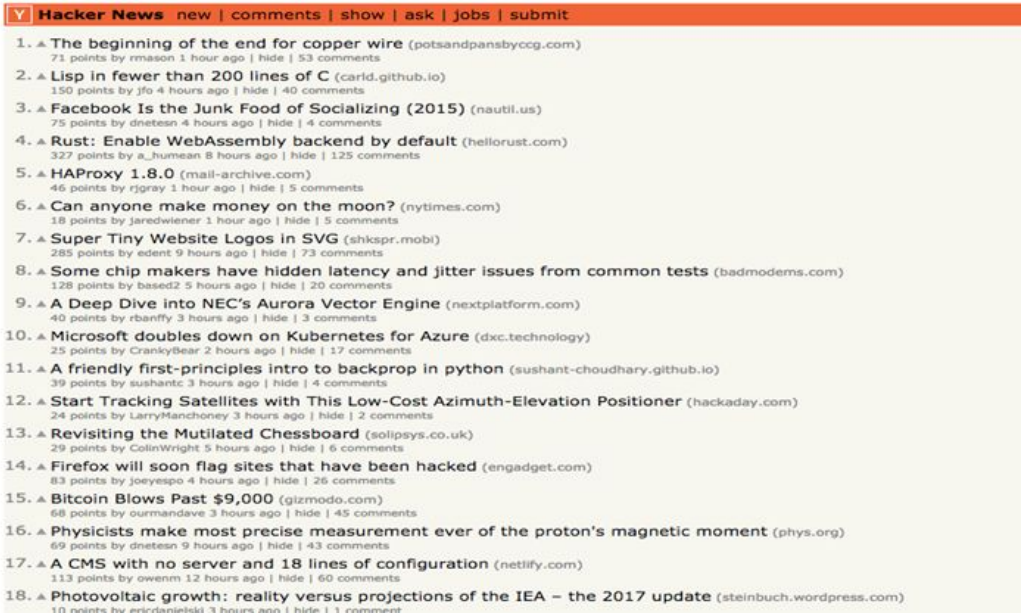
How is our application different?

- 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

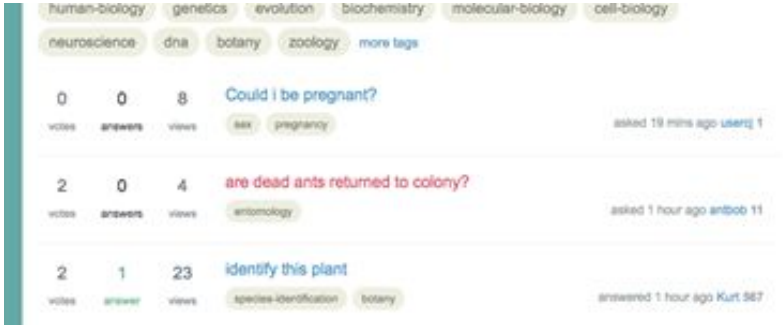
Hackernews

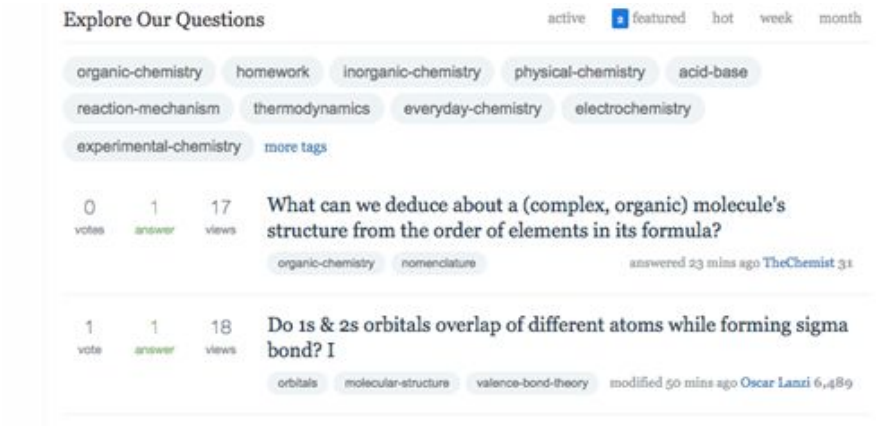
Hacker News is a social news website focusing on computer science and start-ups. The intention was to recreate a community similar to the early days of Reddit.

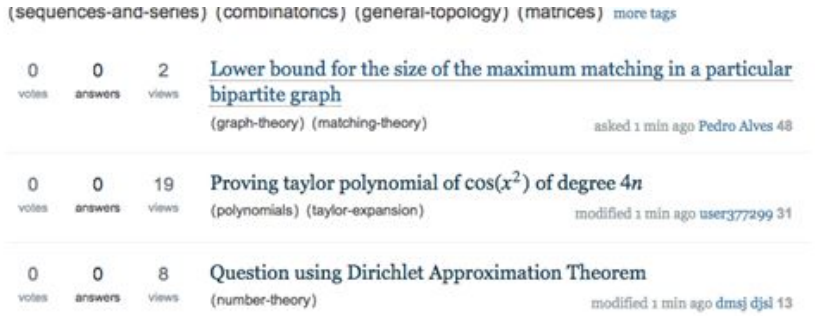


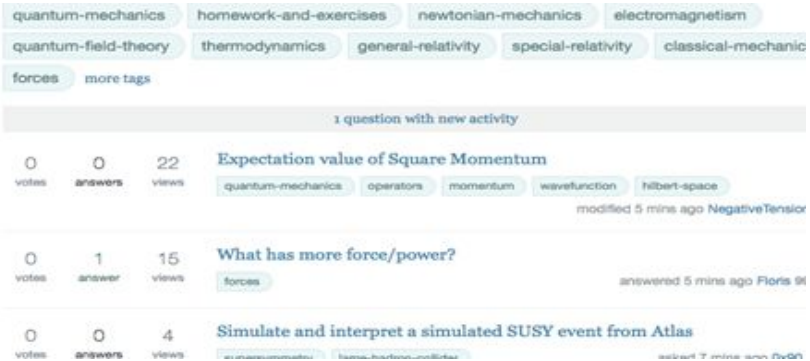
Science Stack Exchange

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









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.

IIF INSTITUT FÜR INFORMATIK FREIBURG 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.

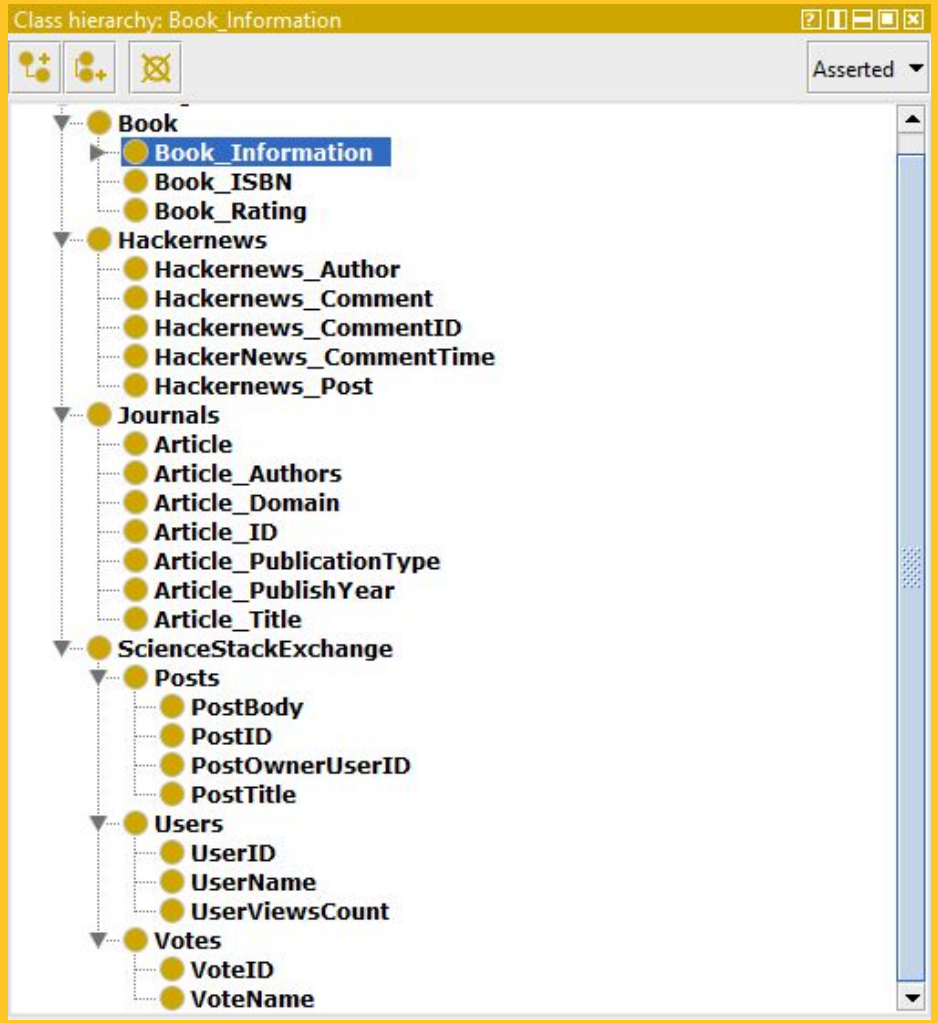
```
- <ref-list>
    + <ref id="B1" content-type="article" article_id="2231.0B1">
    + <ref id="B2" content-type="misc" article_id="2231.0B2">
    + <ref id="B3" content-type="article" article_id="2231.0B3">
    + <ref id="B4" content-type="article" article_id="2231.0B4">
    + <ref id="B5" content-type="article" 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">
```

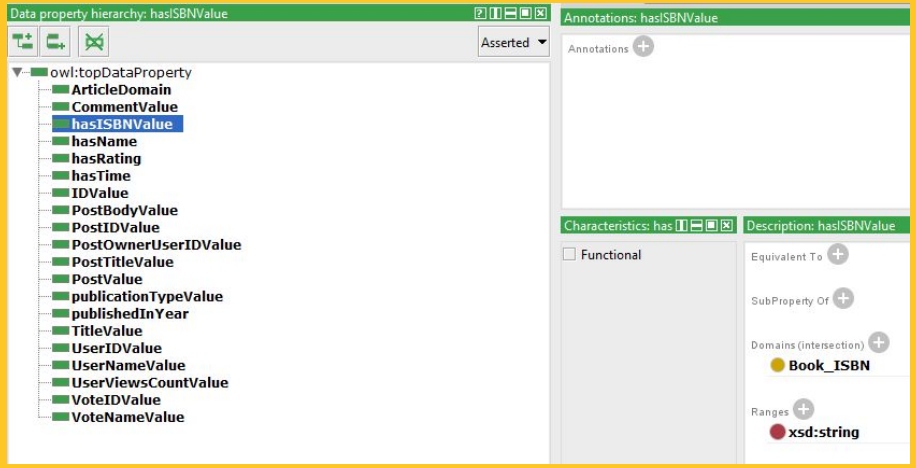
Every XML file has a ref-list under which various articles with unique article_id are mentioned.

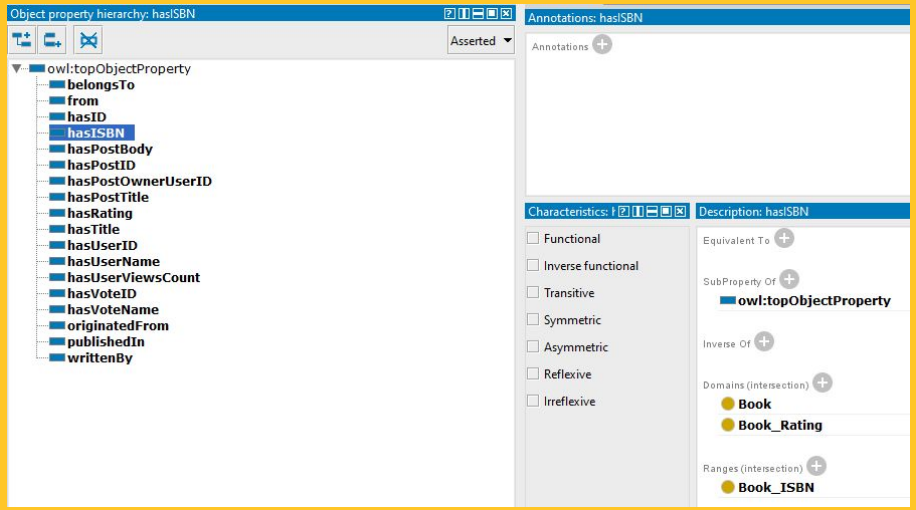
```
- <ref id="B1" content-type="article" article_id="2231.0B1">
  - <nlm-citation publication-type="journal">
    - <person-group person-group-type="author">
      - <name>
          <surname>Steinmann</surname>
          <given-names>P.</given-names>
        </name>
      - <name>
          <surname>Keiser</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.

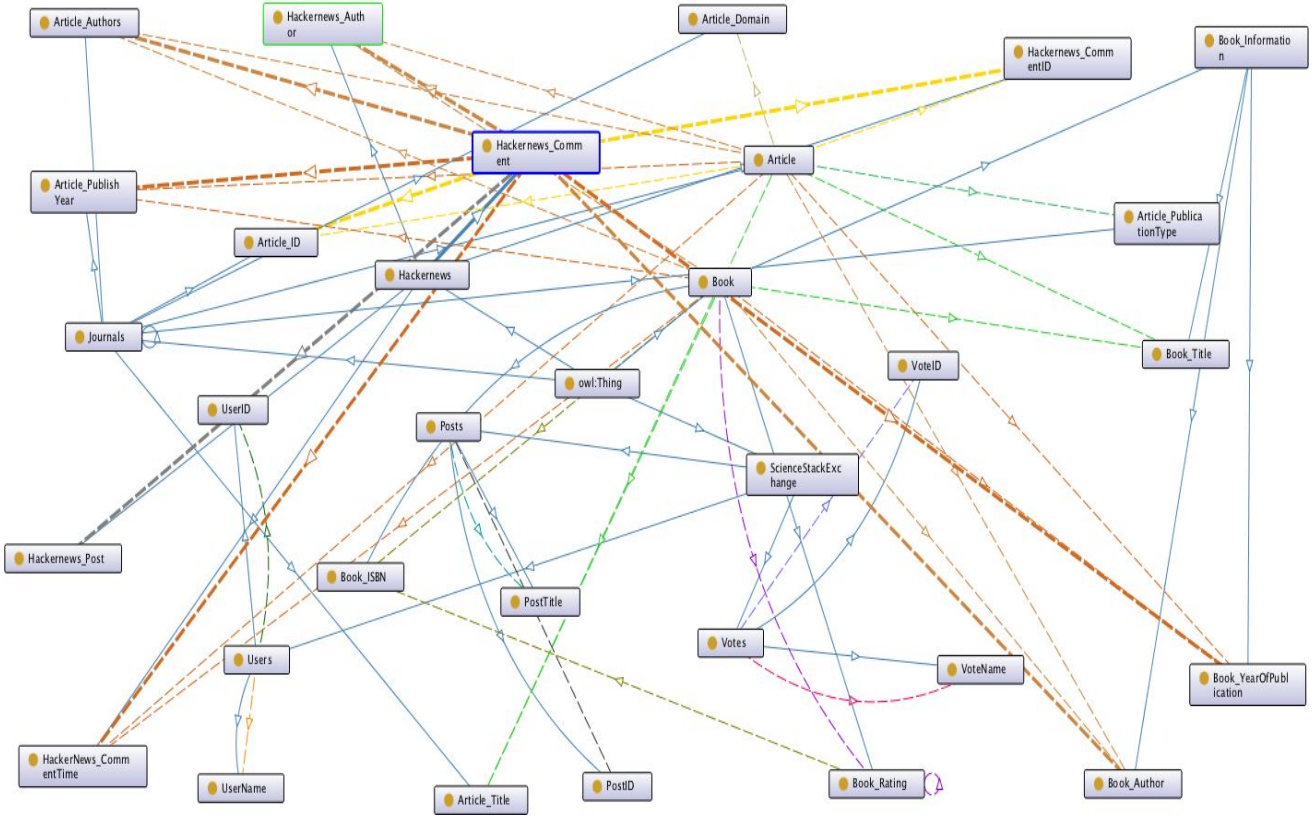
Semantic Data Model


Class hierarchy: Book_Information
Asserted
Book
Book_Information
Book_ISBN
Book_Rating
Hackernews
Hackernews_Author
Hackernews_Comment
Hackernews_CommentID
HackerNews_CommentTime
Hackernews_Post
Journals
Article
Article_Authors
Article_Domain
Article_ID
Article_PublicationType
Article_PublishYear
Article_Title
ScienceStackExchange
Posts
PostBody
PostID
PostOwnerUserID
PostTitle
Users
UserID
UserName
UserViewsCount
Votes
VoteID
VoteName


Data property hierarchy: hasISBNValue
Asserted
owl:topDataProperty
ArticleDomain
CommentValue
hasISBNValue
hasName
hasRating
hasTime
IDValue
PostBodyValue
PostIDValue
PostOwnerUserIDValue
PostTitleValue
PostValue
publicationTypeValue
publishedInYear
TitleValue
UserIDValue
UserNameValue
UserViewsCountValue
VoteIDValue
VoteNameValue
Annotations: hasISBNValue
Annotations
Characteristics: has
Functional
Description: hasISBNValue
Equivalent To
SubProperty Of
Domains (intersection)
Book_ISBN
Ranges
xsd:string


Object property hierarchy: hasISBN
Asserted
owl:topObjectProperty
belongsTo
from
hasID
hasISBN
hasPostBody
hasPostID
hasPostOwnerUserID
hasPostTitle
hasRating
hasTitle
hasUserID
hasUserName
hasUserViewsCount
hasVoteID
hasVoteName
originatedFrom
publishedIn
writtenBy
Annotations: hasISBN
Annotations
Characteristics: h
Functional
Inverse functional
Transitive
Symmetric
Asymmetric
Reflexive
Irreflexive
Description: hasISBN
Equivalent To
SubProperty Of
owl:topObjectProperty
Inverse Of
Domains (intersection)
Book
Book_Rating
Ranges (intersection)
Book_ISBN

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)
- hasVoteID (Domain>Range)
- hasVoteName (Domain>Range)
- originatedFrom (Domain>Range)
- publishedIn (Domain>Range)
- writtenBy (Domain>Range)

Process of Linked Data Generation

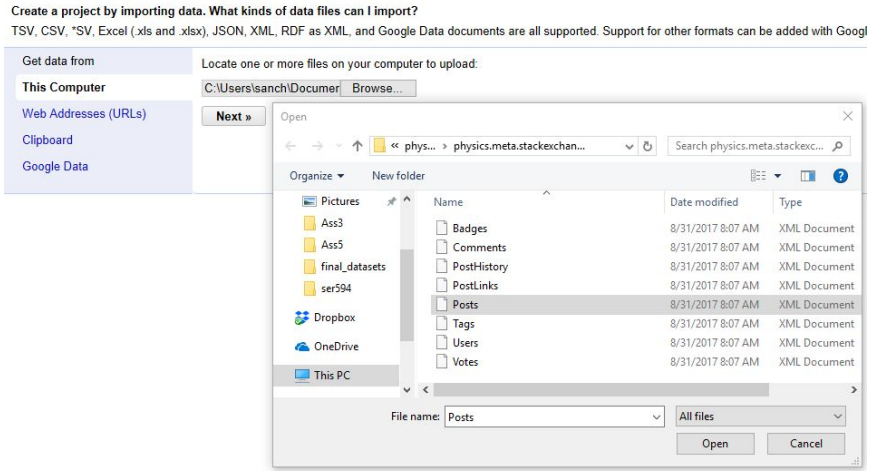
Hackernews

- 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.

	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

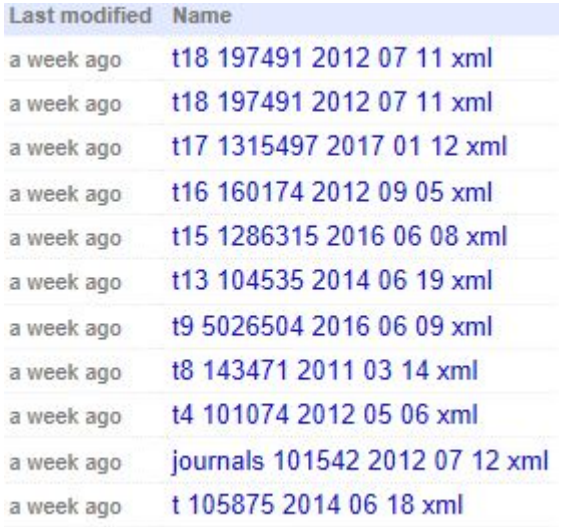
```
@prefix sse: <> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .


<http://127.0.0.1:3333/0> a <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#Book> ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#hasISBN> "195153448" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#hasTitle> "Classical Mythology" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#writtenBy> "Mark P. O. Morford" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#publishedIn> "2002" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#publishedby> "Oxford University Press" .

<http://127.0.0.1:3333/1> a <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#Book> ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#hasISBN> "2005018" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#hasTitle> "Clara Callan" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#writtenBy> "Richard Bruce Wright" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#publishedIn> "2001" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#publishedby> "HarperFlamingo Canada" .

<http://127.0.0.1:3333/2> a <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#Book> ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#hasISBN> "60973129" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#hasTitle> "Decision in Normandy" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#writtenBy> "Carlo D'Este" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-ontology-6#publishedIn> "1991" ;
    <http://www.semanticweb.org/mahimagupta/ontologies/2017/9/untitled-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

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