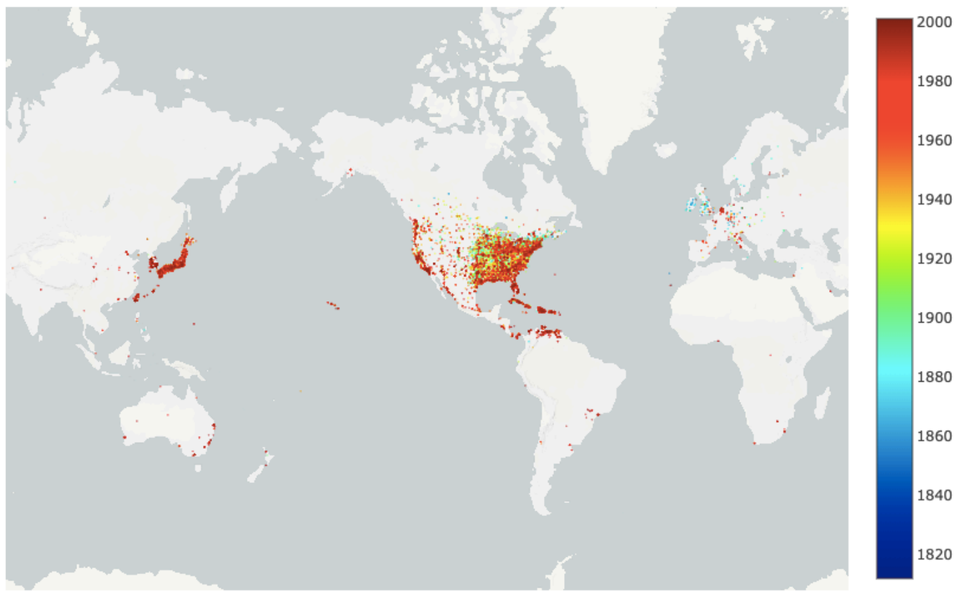
Using Knowledge Graphs to Enhance Sports Analytics

Katherine Schroeder, Marymount University,
Jeremy Abramson, USC information Sciences Institute

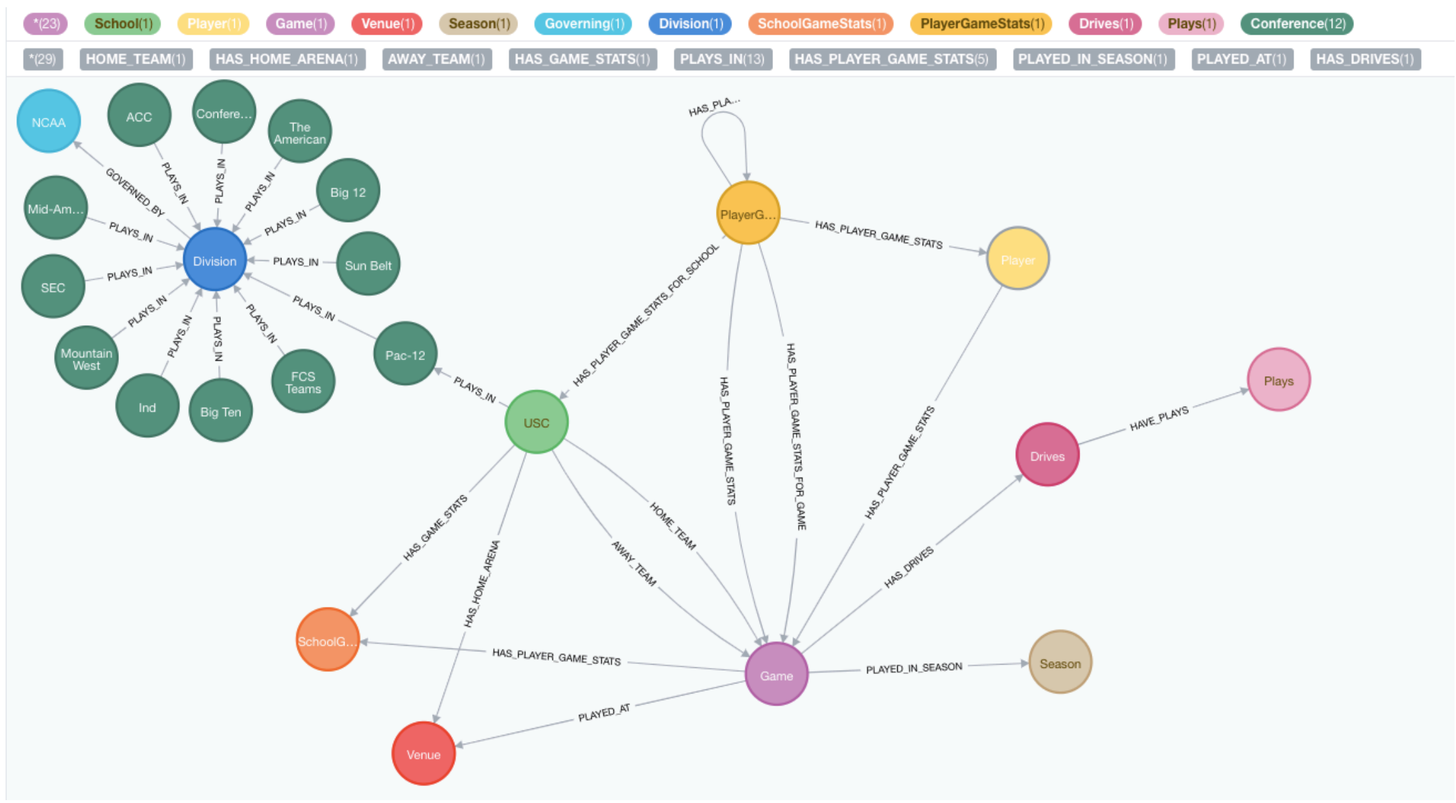
Querying Wikidata

- Using SPARQL Wikidata was queried finding answers to questions like “When were baseball players born and when?” and “What causes of death do baseball players have?”
- Visualizations of the data were made in Python



Building a Knowledge Graph in Neo4j

- Use of data from CollegeFootballData.com for import into the Neo4j graph management system.
- Built Schema in Neo4j.
- Data converted and cleaned.
- Data Fields Imported Include:
  - 11 Conferences
  - 331 Venues
  - 1,664Teams
  - 5,829 Games
  - 129,308 Drives
  - 909,886 Plays



Adding a Property to Wikidata

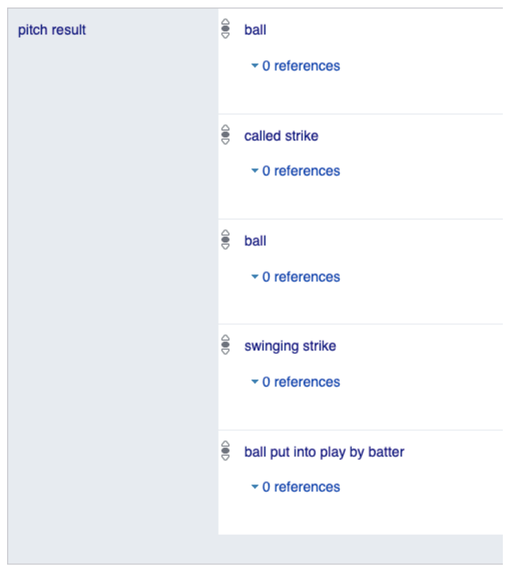
- Identified that Retrosheet IDs, an identifier from a major sabermetrics database, weren’t present for 19,618 baseball players.
- Proposed the property and waited for its merits to be discussed by members of the community.
- Proposed property was added by admin.
- Cleaned data to format it for upload to Wikidata, matching players with existing entries and adding new players who weren’t already represented.
- Upload data to Wikidata.

Play-by-Play Wikibase

- Built schema for Baseball Play-by-Play in local Wikibase instance from Retrosheet data
  - Season
  - Game
  - Innings
  - Player at Bat
  - Pitch Result
  - Events
  - Players

Future Work

- Build a Wikibase showing full Retrosheet play-by-play history, starting with 1921.
  - Local Wikibase Instance transferred to server
  - Data to be cleaned
  - Data to be separated into plays and given identifiers
  - Script written to assign values according to Retrosheet key
  - Using Quickstatements data can be pushed to Wikibase creating a play-by play history



If interested contact Katherine Schroeder - kbs05788@marymount.edu
Work performed under REU Site program
supported by NSF grant #1659886

USC Viterbi School of Engineering Information Sciences Institute