How Much Statistical Data can be Recovered from Alabama Football History?

Steven L. MacCall – University of Alabama

Huapu Liu – University of Alabama

Melissa Anderson – University of Alabama

Deposited 02/04/2020

Citation of Conference Presentation:

MacCall, S., Liu, H., Anderson, M. (2019): How Much Statistical Data can be Recovered from Alabama Football History? *University of Alabama Digital Humanities Conference Tuscaloosa, Alabama 2019.*

How Much Statistical Data Can Be Recovered from Alabama Football History?

Piloting a Crowdsourced Approach Using Wikibase as Data Repository

Steven L. MacCall, PhD
Huapu Liu, MLIS
Melissa Anderson, SLIS Graduate Student

School of Library and Information Studies
The University of Alabama

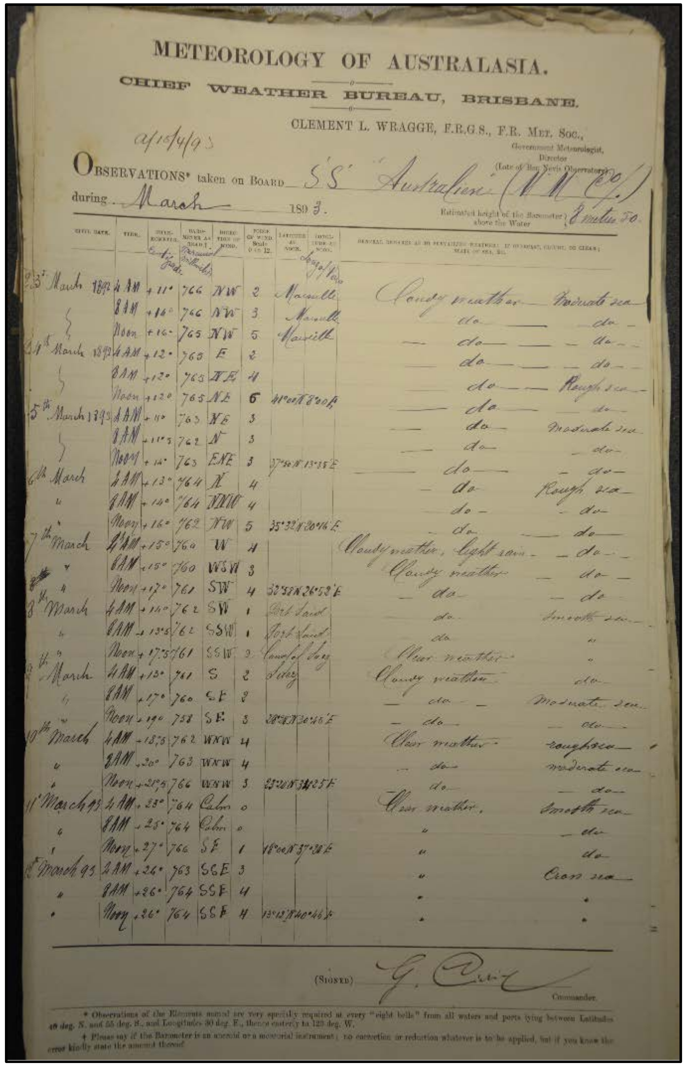
Agenda

- Setting the stage: The weather data rescue example.
- Background: Discussions with Ken Gaddy of the Paul W. Bryant Museum:
  - Overabundant digital content and the need for a player database
  - Experimenting with play-by-play datasets to index plays of games (play database)
- Data-driven indexing to create a “play database”:
  - Accomplished: Transforming JSON formatted born digital play-by-play data
  - Our current study: What about data rescue from print formats?
- Our data processing pipeline for recovering historical data:
  - 1992 games: Transcribing and data extraction method
  - 1961 games: Data extraction method
- Results and conclusions.
- Postscript: How should we label what we are doing?

Setting the Stage: Rescuing Weather Data

- Efforts underway to recover old weather observations extracted from ship logbooks.

- Structured data suitable for extraction to spreadsheets for integration and further analysis.



Background for our Study

- For over a decade, PI has had discussions with Ken Gaddy of the Paul W. Bryant Museum:
  - Overabundant historical and current multimedia content
  - The need for a player database
- During this time, there was a realization that indexing individual images was not realistic because of too much multimedia content.
- The answer seems to lie with using play-by-play datasets to index plays of games (play database) rather than each image or video clip.
- So far, we have accomplished the semantic indexing of all plays from the 2017 Alabama football season using Wikibase as our structured data repository.

Wikibase as a Structured Data Repository

- Wikibase is open-source software of the WikiMedia Foundation:
  - Drives their Wikidata service
  - Wikibase Client is a MediaWiki extension that can turn a MediaWiki installation into a client of a structured data repository
  - Wikibase is beginning to be deployed in LIS institutions such as OCLC's Project Passage
- In fall 2018, we started using a local instance of Wikibase to accomplish our semantic indexing of the 2017 season:
  - RDF triples describe attributes of plays (example play)
  - SPARQL endpoint allows for querying of play database (sample query)

Our Data Processing Pipeline – Overview

- Recovering data from sources:
  - Transcribe and extract paper-based play-by-play data (1992 games)
  - Extract paper-based play-by-play data from non-structured textual data (1961 games)

- Data wrangling of recovered data:
  - Formatted data in spreadsheets according to relevant data type from our Wikibase properties list (e.g., down and distance; game score at start of play…)
  - Entity reconciliation using custom Python script: Mapped certain properties' values to Wikibase Q numbers (e.g., player names, type of play…)

- Batch upload of data to Wikibase using QuickStatements tool.

Data Sources for this Study

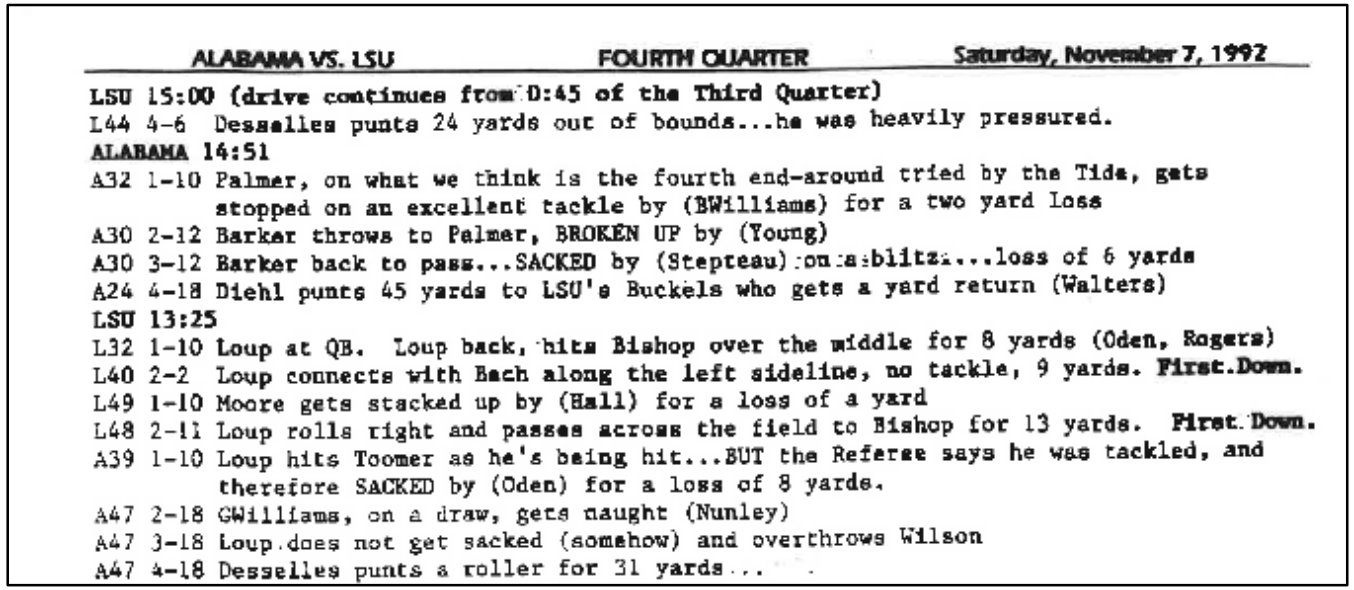
- 1992 data sources:
  - Two games from 1992 season
    - Alabama versus LA Tech on September 26, 1992 at Legion Field in Birmingham
    - Alabama versus LSU on November 7, 1992 at Tiger Stadium in Baton Rouge, LA
  - Typewritten datasets on paper
  - Play-by-play game logs and drive charts recorded by UA Athletics officials
  - Preserved at the Paul W. Bryant Museum's research documents collection

- 1961 data sources:
  - Two games from 1961 season
    - Alabama versus Tulane on September 30, 1961 at Ladd Stadium in Mobile, AL
    - Alabama versus Vanderbilt on October 7, 1961 at Dudley Field in Nashville, TN
  - Text from primary Tuscaloosa News article for each game
  - Accessible online at Google newspaper archive

Our Zooniverse Process

(1992 Dataset)

- Designed transcription and data extraction tasks with written instructions and video tutorials.
- Conducted usability testing.
- Recruited volunteers.
- Exported .csv files datasets.
- Developed data wrangling and cleaning workflows.



ALABAMA VS. LSU FOURTH QUARTER Saturday, November 7, 1992

LSU 15:00 (drive continues from 0:45 of the Third Quarter)
L44 4-6 Desselles punts 24 yards out of bounds...he was heavily pressured.
ALABAMA 14:51
A32 1-10 Palmer, on what we think is the fourth end-around tried by the Tide, gets stopped on an excellent tackle by (BWilliams) for a two yard loss
A30 2-12 Barker throws to Palmer, BROKEN UP by (Young)
A30 3-12 Barker back to pass...SACKED by (Stepteau) on a blitz....loss of 6 yards
A24 4-18 Diehl punts 45 yards to LSU's Buckels who gets a yard return (Walters)
LSU 13:25
L32 1-10 Loup at QB. Loup back, hits Bishop over the middle for 8 yards (Oden, Rogers)
L40 2-2 Loup connects with Bach along the left sideline, no tackle, 9 yards. **First.Down.**
L49 1-10 Moore gets stacked up by (Hall) for a loss of a yard
L48 2-11 Loup rolls right and passes across the field to Bishop for 13 yards. **First.Down.**
A39 1-10 Loup hits Toomer as he's being hit...BUT the Referee says he was tackled, and therefore SACKED by (Oden) for a loss of 8 yards.
A47 2-18 GWilliams, on a draw, gets naught (Nunley)
A47 3-18 Loup does not get sacked (somehow) and overthrows Wilson
A47 4-18 Desselles punts a roller for 31 yards...

Post-Download Data Processing

(1992 Dataset)

- Challenges presented by Zooniverse formatting of downloaded data:
  - The exported data is in the format of text instead of tabular data
  - Data is separated into different workflows (Play-by-Play Transcription; Play Type; Resulting Yardage; Offense Team; Featured Players)

- Data Manipulation:
  - Transforming the data into needed format (Text to columns by delimiters; Transpose columns to rows/ rows to columns)
  - Connecting data from different tasks into one spreadsheet (Matching the data entry from different workflows)

1992 Data Recovery Summary

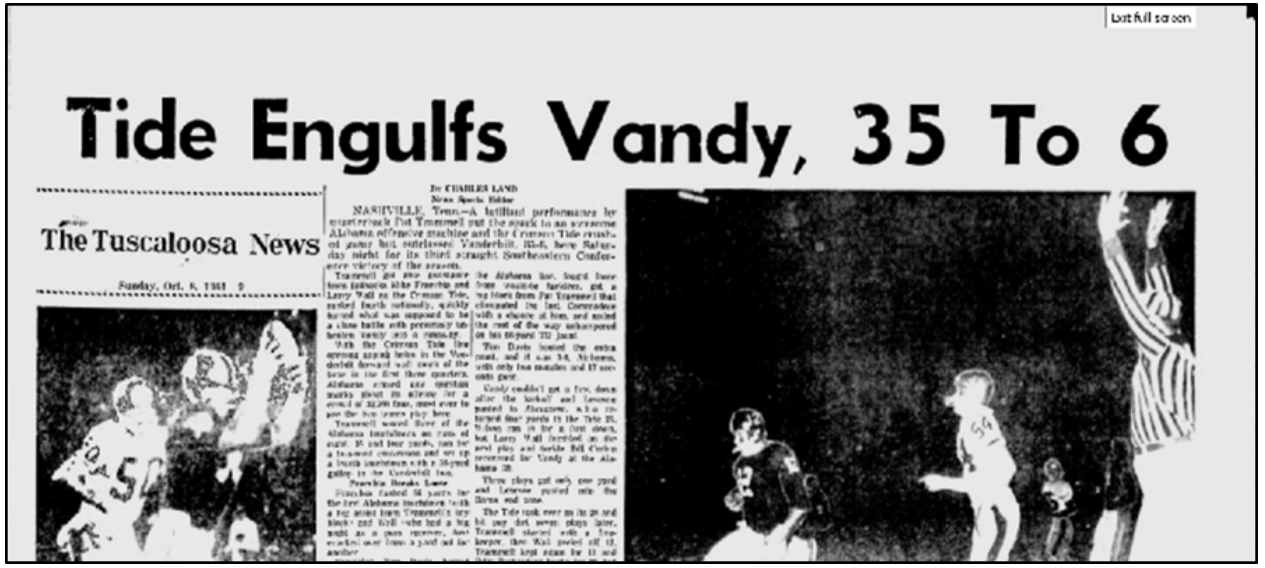
- The use of volunteers on the Zooniverse Platform was workable.
- There are challenges to scaling up:
  - Preprocessing, cleaning, and volunteer management are time consuming
  - There are over 100 games with typewritten, paper datasources across all seasons of UA football
- More advanced OCR might reduce role of crowdsourcing.

Extracting Play Data from Newspaper Articles
(1961 Dataset)

Close reading of textual sources

- Extracted play-by-play data read from text to spreadsheet:
  - Articles sectioned into quarters from game
  - Player name data
  - Numerical data (e.g., play yardage and/or scoring outcomes)
- Recorded inferences sanctioned by data (e.g., drive data).

1961 example data source



Tide Engulfs Vandy, 35 To 6

The Tuscaloosa News

Sunday, Oct. 8, 1961

Study Results and Conclusion

- How much data did we recover?
  - Results for 1992:
    - Alabama versus LA Tech: Data on **167 play actions** and **28 drives**
    - Alabama versus LSU: Data on **185 play actions** and **26 drives**
  - Results for 1961:
    - Alabama versus Tulane: Data on **104 play actions** and **22 drives**
    - Alabama versus Vanderbilt: Data on **129 play actions** and **25 drives**
- Valuable data is available to reconstruct the past at a more granular level (play-by-play) than is currently the case.
- For scaling, our process will need “machine-aided” assistance, including: intelligent OCR, Named Entity Recognition, and digital image processing.

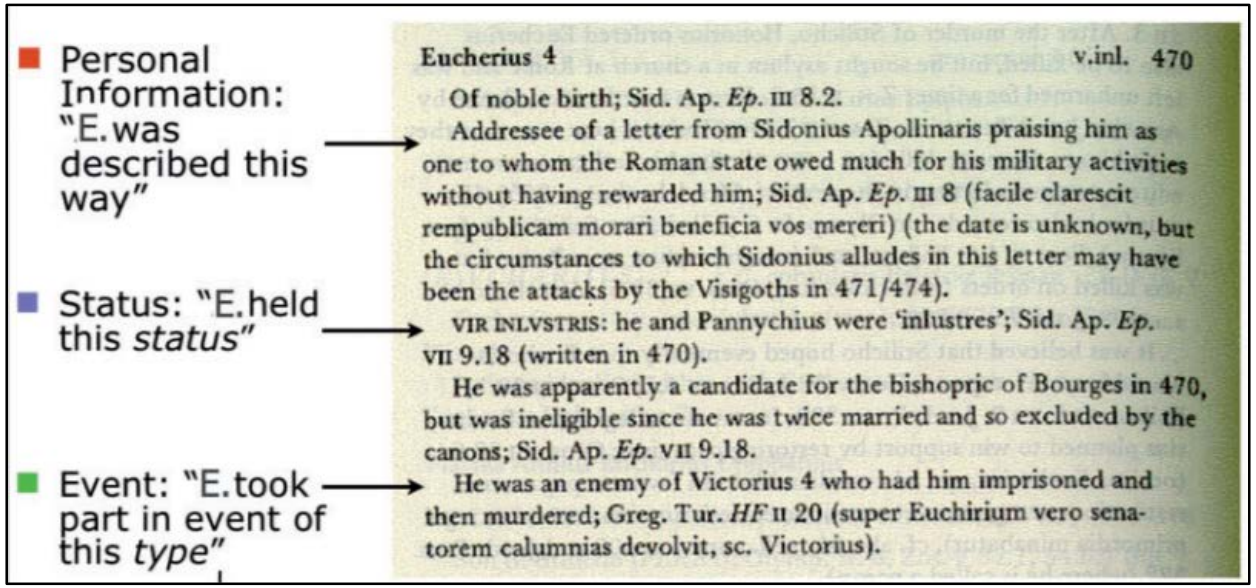
Postscript: How Should We Label This Work?

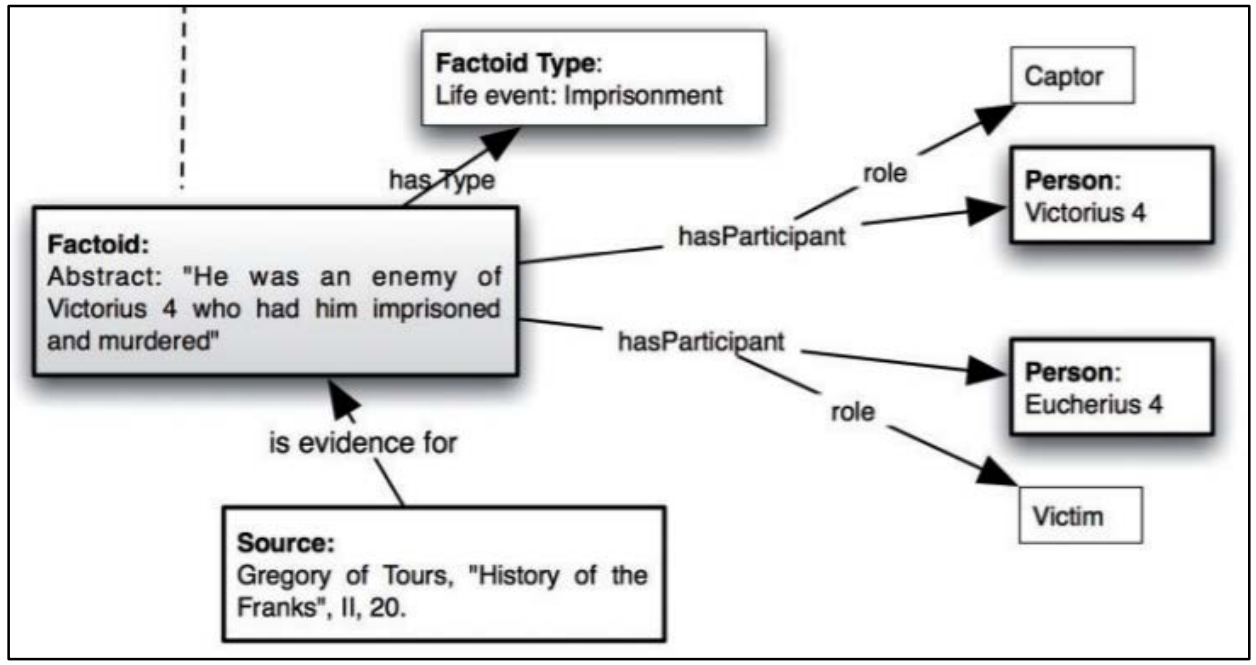
- How should our “player database” work be labeled from a DH perspective?
- Factoid Prosopographical Database aligns with long practice:
  - Prosopography is a historical research methodology that investigates the common characteristics of a group of people to create a collective biography (Stone, 1971).
  - Derives from *prosopon*, meaning “person.”
  - Factoid prosopography is a structured data approach that links people to information about them via spots in primary sources (Pasin & Bradley, 2015).
  - Recent projects explore approaches to making prosopographical databases available on the Semantic Web (Pasin & Bradley, 2015).

Textual vs. Structural

- Prosopographical reference tools were originally historical prose.

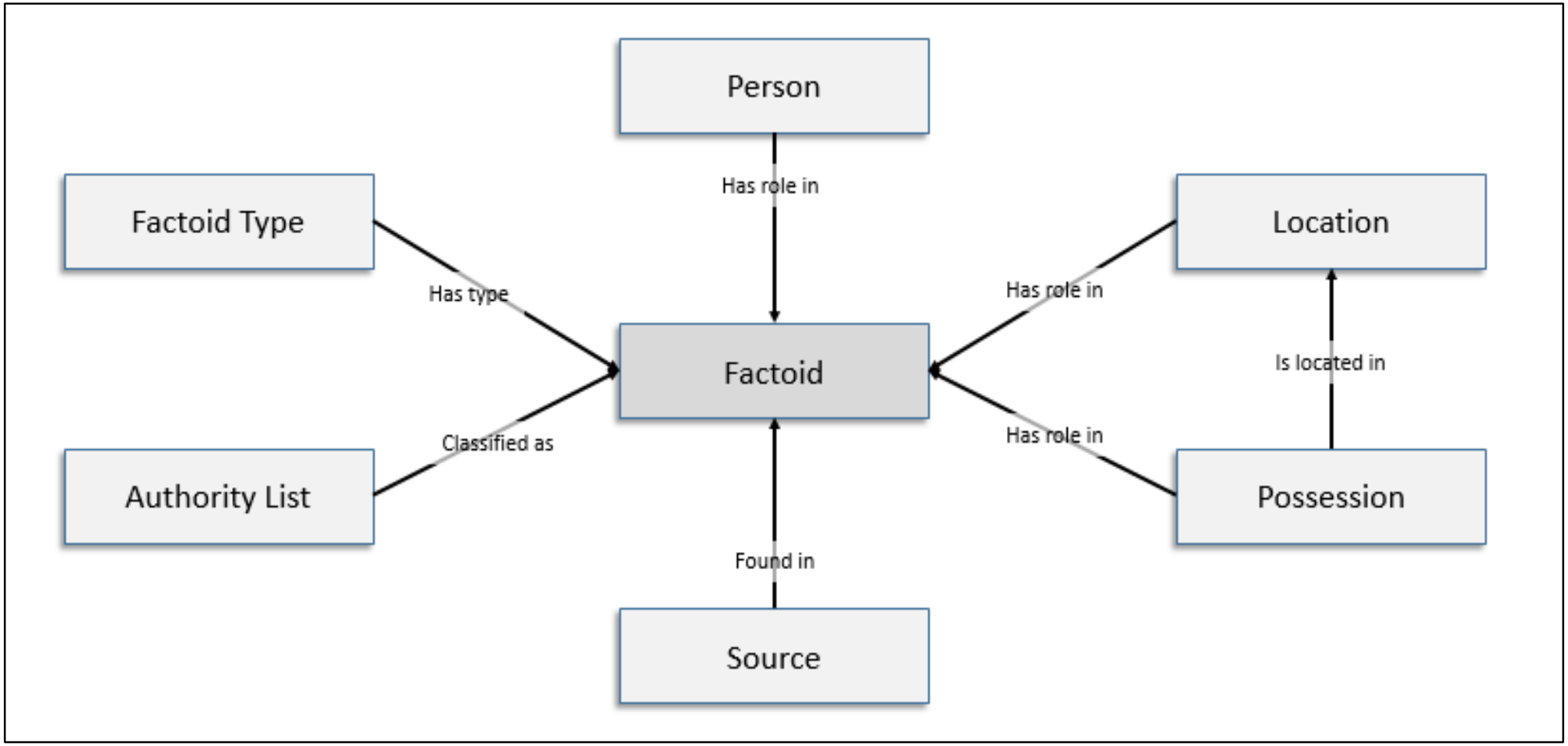
- In the 1990s, King's College London developed the factoid model of prosopographical databases.





[Michele Pasin and John Bradley, 2015]

FACTOID = a spot in a source that acts as a structural nexus connecting together historical sources, people, places, possessions, personal relationships, titles, etc.



[Pasin and Bradley, 2015]

Final Thoughts: Toward Agonography?

- How should our "*play* database" work be labeled from a DH perspective?

- If prosopography is about people, are we rather creating the first sports-related agonography?
    - *Agon* can be translated as "competition," "competition at games," or "contest" (*OED;* Merriam-Webster).
    - May provide terminology to describe play database.
    - Has been used to refer to electronic game design but not formally associated with any current resource or practice.

Thanks to Our Volunteers!

- Dr. James Elmborg, Director of the School of Library and Information Studies (SLIS)
- Dr. Elizabeth Aversa, President of the Osher Lifelong Learning Institute, UA
- Mr. Rocco Aversa, Tuscaloosa
- Dr. Ann Bourne, SLIS Assistant Director
- Dr. Steven Yates, SLIS Assistant Director & School Library Media Coordinator
- Jonathan Bowen, Graduate Student, UA School of Public Administration
- Emily Mayers-Twist, JD, SLIS Graduate Student
- Dione Thrift, Reference Assistant, North Shelby Library, SLIS Graduate Student
- Emma Davis, Library Specialist, Hoover Public Library, SLIS Graduate Student
- Brandon Wicks, MFA, Visiting Assistant Professor of Writing, Emory University
- Dr. Walter Ward, Associate Professor of History, UAB