Generating metadata subject labels with Doc2Vec and DBPedia

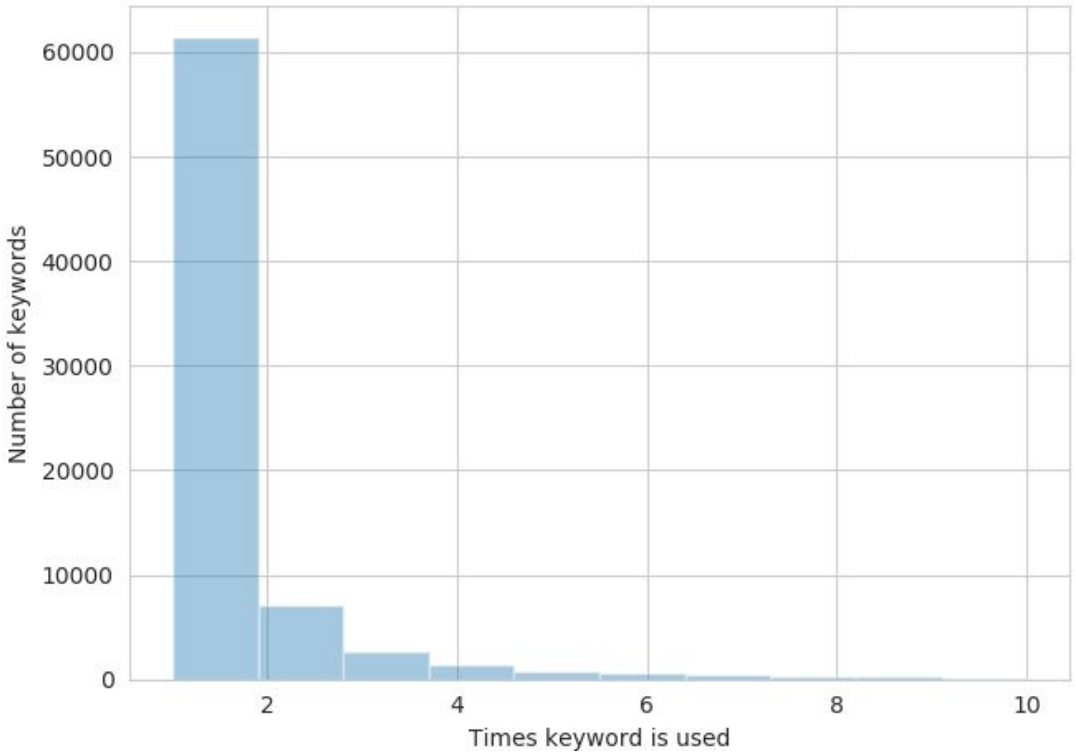
Charlie Harper
Digital Scholarship Specialist
KELVIN SMITH LIBRARY
CASE WESTERN RESERVE UNIVERSITY

Inspiration

---

1) Constant discussion of topic modeling as “solution” to improving data discovery [useful internally]
2) General issue of document clustering without existing cluster labels and problem of cluster interpretability
3) ETDs are extremely hard to search, but contain important information on trends in fields and not yet published data
4) Author-assigned keywords for ETDs are problematic

Author-assigned keywords for Ohio ETDs (2015-2020)



Around 77k author-assigned keywords

80% of terms are only used once

Others are too common/generic

subject	
psychology	2038
education	1589
biology	1324
chemistry	1217
mechanical engineering	1192

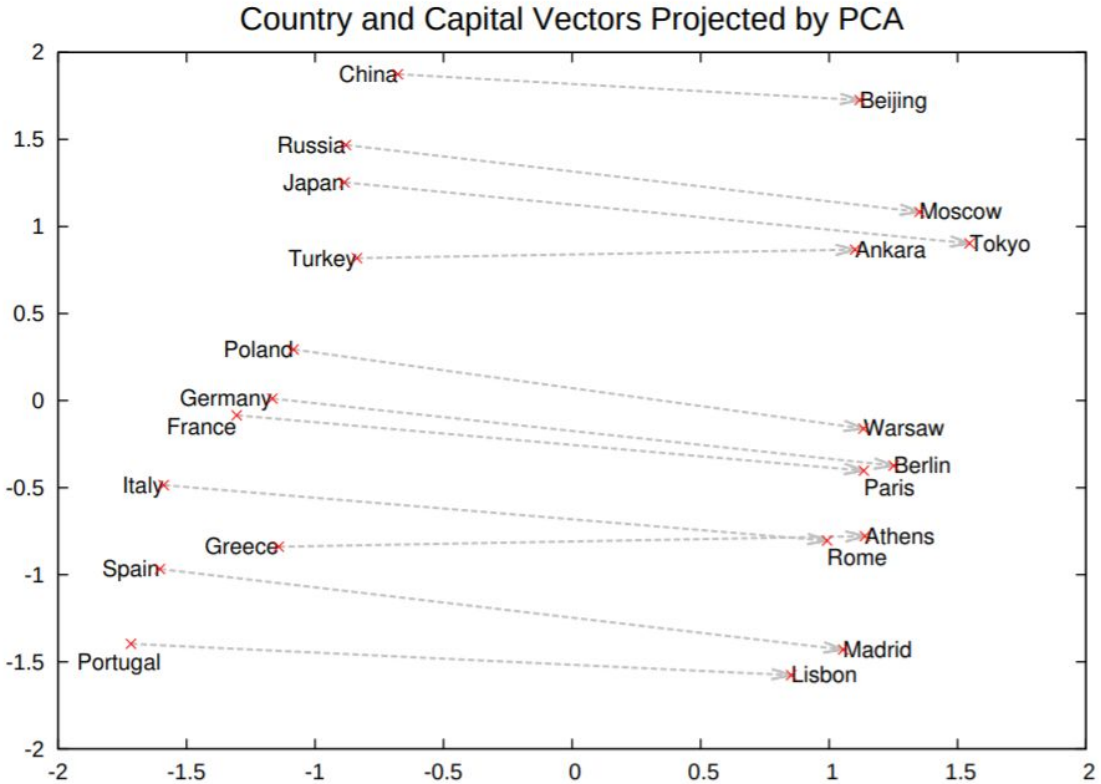
Model Choice: Doc2Vec

Word2Vec maps words into high-dimension vector space

Unsupervised approach

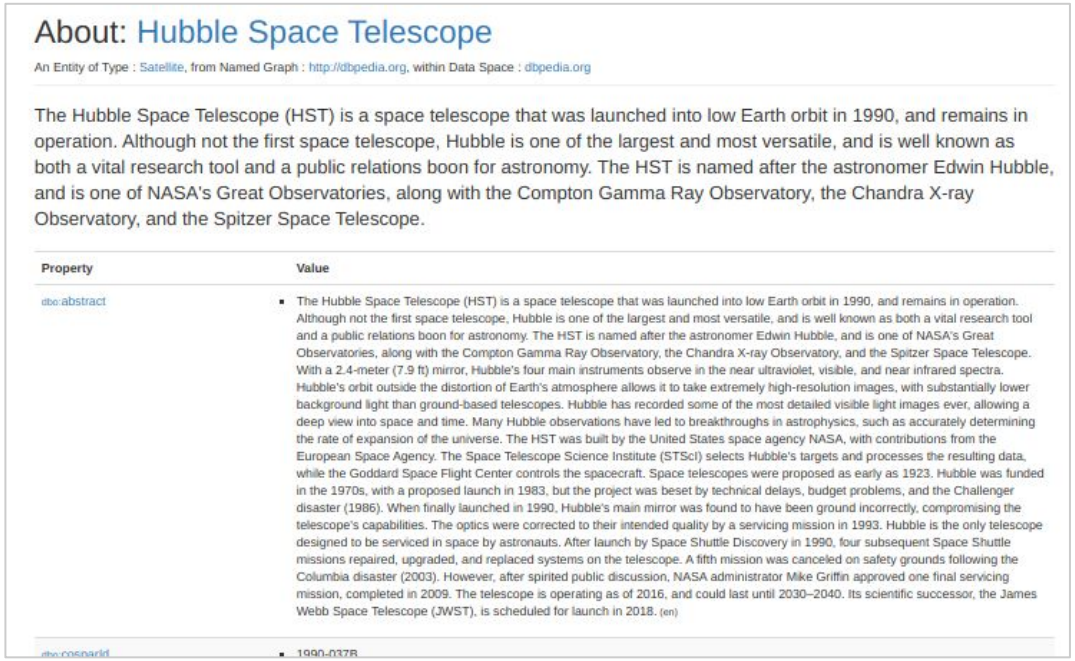
Captures important relationships

Doc2Vec extends to full texts (Gensim Library in Python)



(From Mikolov et al. 2013. https://arxiv.org/abs/1310.4546)

Training Data: DBPedia



About: Hubble Space Telescope

An Entity of Type : Satellite, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

The Hubble Space Telescope (HST) is a space telescope that was launched into low Earth orbit in 1990, and remains in operation. Although not the first space telescope, Hubble is one of the largest and most versatile, and is well known as both a vital research tool and a public relations boon for astronomy. The HST is named after the astronomer Edwin Hubble, and is one of NASA's Great Observatories, along with the Compton Gamma Ray Observatory, the Chandra X-ray Observatory, and the Spitzer Space Telescope.

Property	Value
dbo:abstract	The Hubble Space Telescope (HST) is a space telescope that was launched into low Earth orbit in 1990, and remains in operation. Although not the first space telescope, Hubble is one of the largest and most versatile, and is well known as both a vital research tool and a public relations boon for astronomy. The HST is named after the astronomer Edwin Hubble, and is one of NASA's Great Observatories, along with the Compton Gamma Ray Observatory, the Chandra X-ray Observatory, and the Spitzer Space Telescope. With a 2.4-meter (7.9 ft) mirror, Hubble's four main instruments observe in the near ultraviolet, visible, and near infrared spectra. Hubble's orbit outside the distortion of Earth's atmosphere allows it to take extremely high-resolution images, with substantially lower background light than ground-based telescopes. Hubble has recorded some of the most detailed visible light images ever, allowing a deep view into space and time. Many Hubble observations have led to breakthroughs in astrophysics, such as accurately determining the rate of expansion of the universe. The HST was built by the United States space agency NASA, with contributions from the European Space Agency. The Space Telescope Science Institute (STScI) selects Hubble's targets and processes the resulting data, while the Goddard Space Flight Center controls the spacecraft. Space telescopes were proposed as early as 1923. Hubble was funded in the 1970s, with a proposed launch in 1983, but the project was beset by technical delays, budget problems, and the Challenger disaster (1986). When finally launched in 1990, Hubble's main mirror was found to have been ground incorrectly, compromising the telescope's capabilities. The optics were corrected to their intended quality by a servicing mission in 1993. Hubble is the only telescope designed to be serviced in space by astronauts. After launch by Space Shuttle Discovery in 1990, four subsequent Space Shuttle missions repaired, upgraded, and replaced systems on the telescope. A fifth mission was canceled on safety grounds following the Columbia disaster (2003). However, after spirited public discussion, NASA administrator Mike Griffin approved one final servicing mission, completed in 2009. The telescope is operating as of 2016, and could last until 2030–2040. Its scientific successor, the James Webb Space Telescope (JWST), is scheduled for launch in 2018. (en)
dbo:cospar	1990-037B

Where do you find a giant pre-labelled corpus? DBPedia!

Power in open, linked, and multilingual aspects.

Long abstracts of DBPedia version 2016-10 (~5 million pages).

Python scripts on Case Western's High Performance Computing

Artificial “Intelligence”

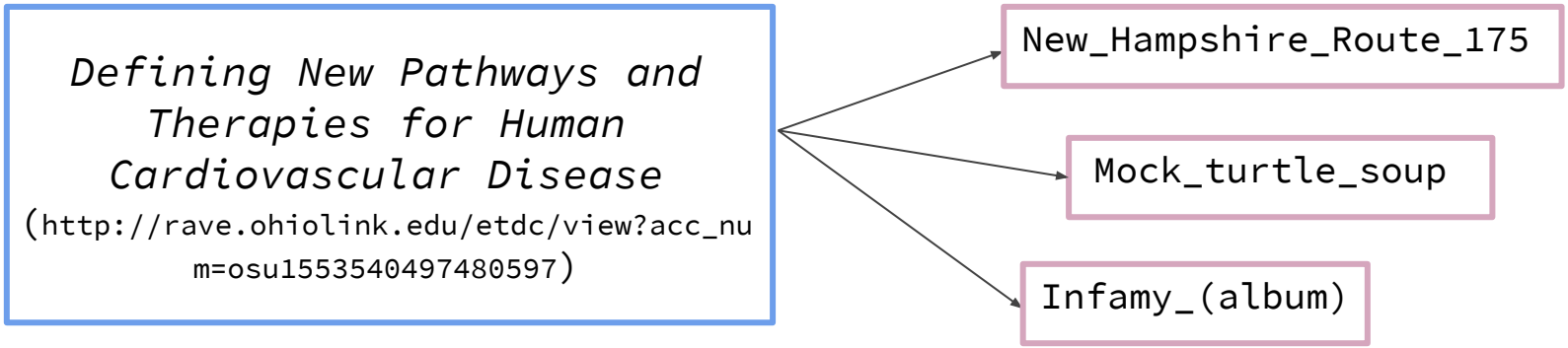
Vectorized an unseen ETD abstract with the trained model Doc2Vec model

Assign the five spatially nearest DBPedia abstracts as your subject tags

Artificial “Intelligence”

Vectorized an unseen ETD abstract with the trained model Doc2Vec model

Assign the five spatially nearest DBPedia abstracts as your subject tags

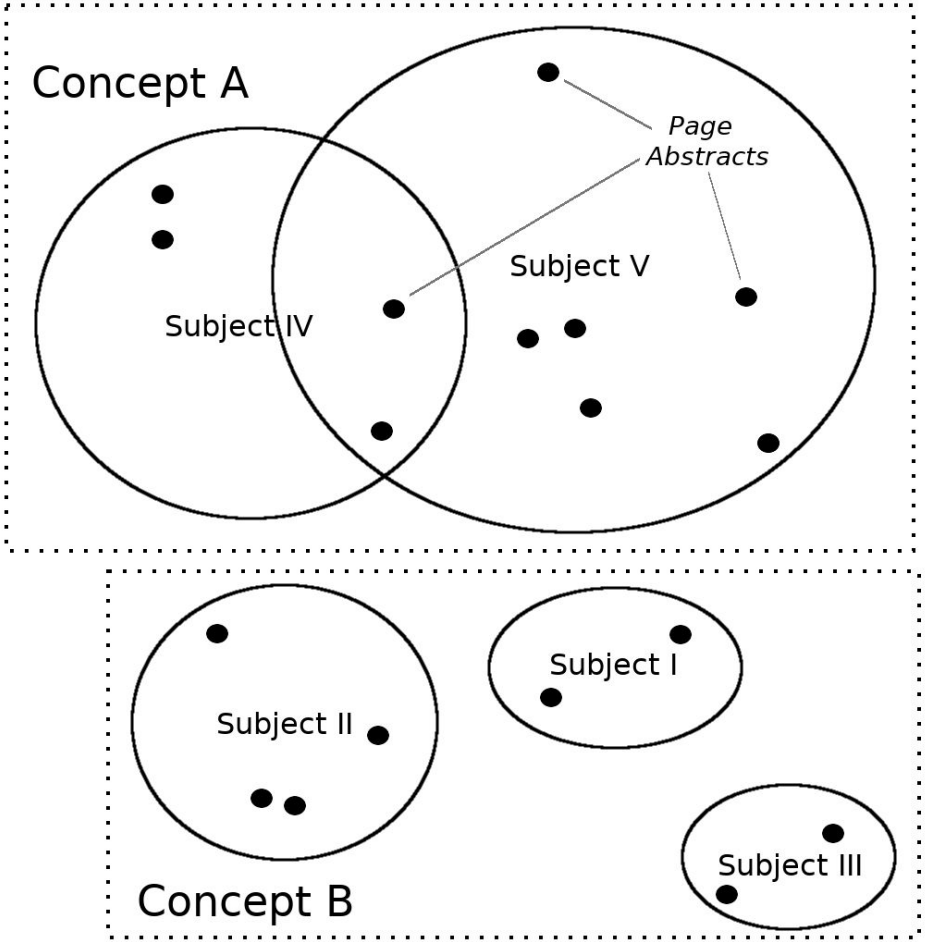


dct:subject and skos:broader

Vector space was overcrowded with pages

Moving up through links would capture higher-level of information and space out vectors

Approach of averaging page vectors to create “subject” and “concept” tags



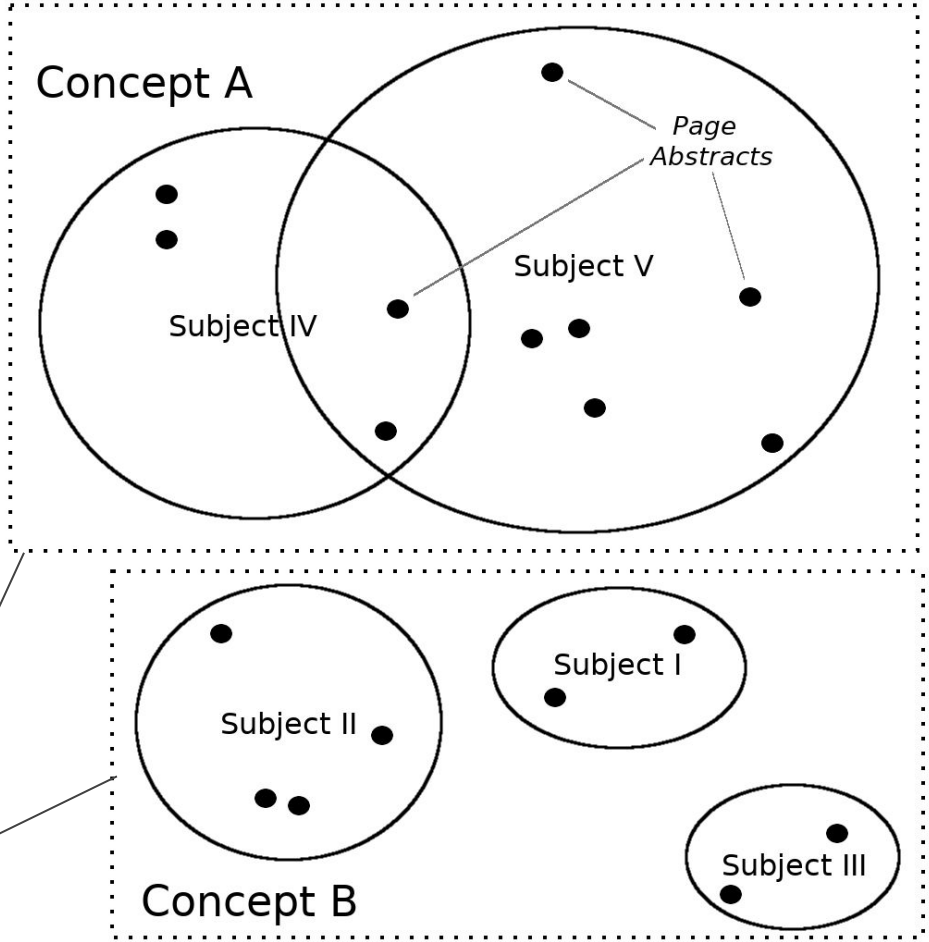
dct:subject and skos:broader

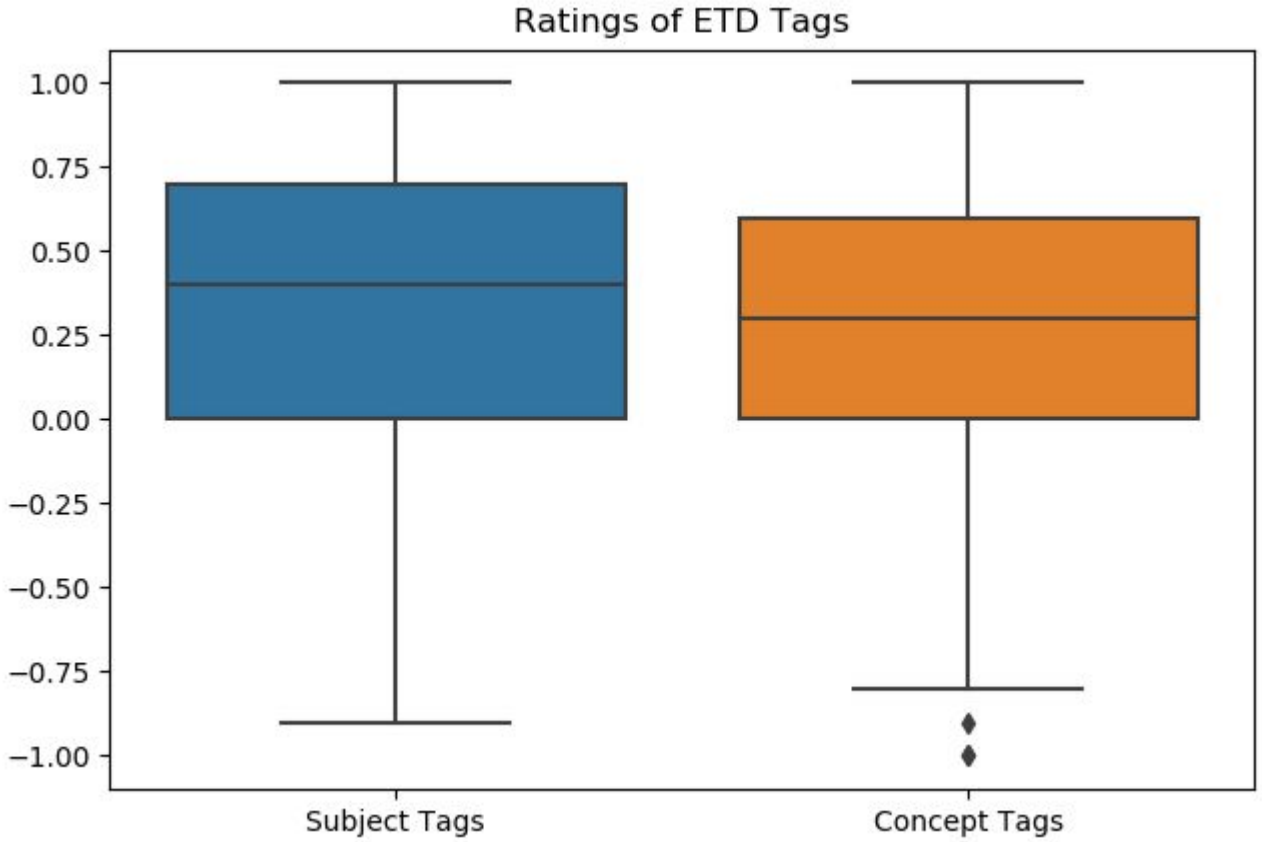
Vector space was overcrowded with pages

Moving up through links would capture higher-level of information and space out vectors

Approach of averaging page vectors to create “subject” and “concept” tags

Tumor_supressor_genes
Programmed_cell_death
Cellular_processes
Transport_proteins





Small sample from 2019 and subset of DBPedia pages

1, 0, -1 rating of assigned tags

On average initial tests show relevant tags

Building Out and Refining

Lots of parameters to explore in the model and decisions about training data to make

Reducing tags further, how to cull?

Expanding ETD dataset (Ohio schools are STEM heavy)

Assessing tag quality / engaging subject experts

Proto-typing visualizations and UIs for searching

Searching Ohio ETD Tags

astronom

- Astronomical X-ray sources
- Astronomical catalogues of galaxy clusters
- Astronomical controversies
- Astronomical databases
- Astronomical events
- Astronomical events of the Solar System
- Astronomical imaging
- Astronomical instruments
- Astronomical objects
- Astronomical objects by year of discovery

Searching Ohio ETD Tags

Astronomical imaging

Identification and Characterization of Long Period Variable Stars in the Globular Cluster M69

Husband, Paul W, Jr.
(2017-08-02, Bowling Green State University / OhioLINK)

Observations of the globular cluster M69 were taken from August 2009 to September 2014 using the 0.4-m PROMPT #4 telescope in Chile. This telescope took observations in V and I bandpass filters, approximately once per week, for ten months of each year. Using the image subtraction software ...

Permalink

Searching Ohio ETD Tags

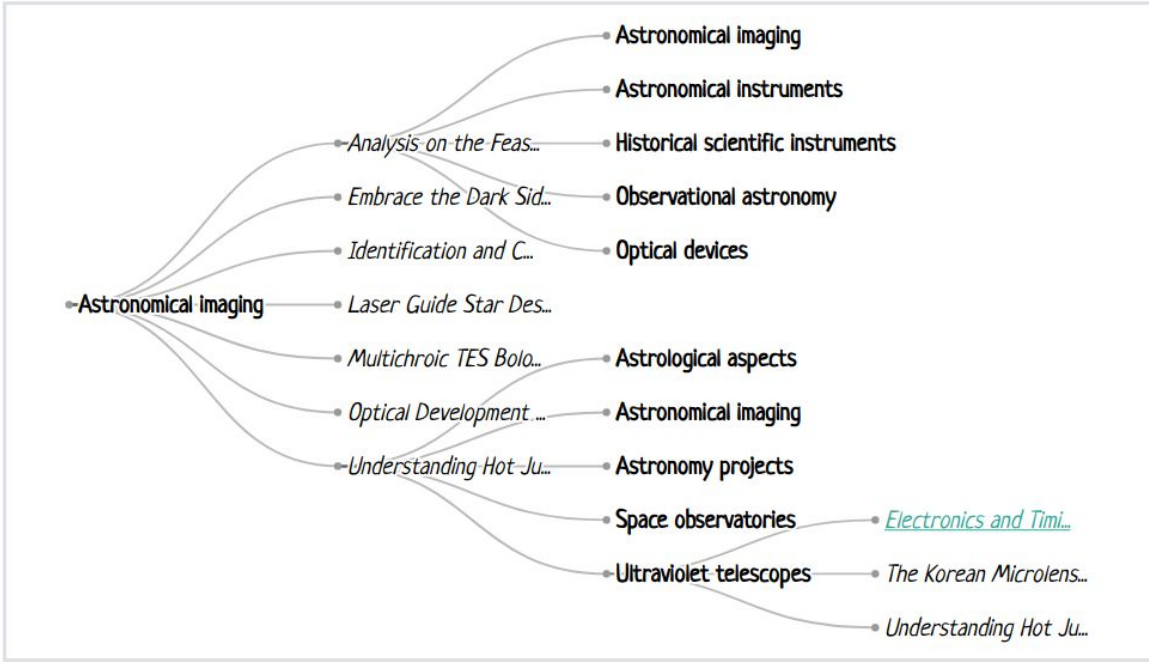
Astronomical imaging

Electronics and Timing for the AugerPrime Upgrade and Correlation of Starburst Galaxies with Arrival Directions of Ultra High Energy Cosmic Rays

Halliday, Robert Paul
(2019-05-23, Case Western Reserve University School of Graduate Studies / OhioLINK)

In this dissertation, we will describe work completed towards the Pierre Auger Observatory's AugerPrime Upgrade as well as auxiliary timing work, hardware design and finally a test of correlations of Starburst Galaxies with the arrival directions of Ultra High Energy Cosmic Rays (UHECRs). In the fir...

Permalink



Acknowledgements

Anne Kummer (Metadata Librarian, CWRU)

Shelby Stuart (Electronic Resources Librarian, CWRU)

Evan Meszaros (Research Services Librarian, CWRU)

This work made use of the High Performance Computing Resource in the Core Facility for Advanced Research Computing at Case Western Reserve University.