

SUPPORTING EARLY MISSION CONCEPT EVALUATION THROUGH
NATURAL LANGUAGE PROCESSING

A Thesis

by

BENJAMIN CADE SIMPSON

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Chair of Committee, Daniel Selva

Committee Members, Bonnie Dunbar
Ruihong Huang

Head of Department, Ivett Leyva

May 2022

Major Subject: Aerospace Engineering

Copyright 2022 Benjamin Cade Simpson

ABSTRACT

Proposal evaluation of pre-Phase A mission concepts is largely based on the input from subject matter experts who determine the scientific merit of a mission concept based on a number of criteria including: the relevance of the mission objectives to national and international priorities; the existence of a complete set of measurement, instrument, and platform requirements that are traceable to the mission objectives; and several others. The Science Traceability Matrix is a standard tool used to articulate this relevance and traceability and therefore is a key input to this reviewing process. However, inconsistencies in the structure and vocabulary used in the Science Traceability Matrix and other sections of the proposal across organizations make this process challenging and time-consuming. At the same time, as part of the Digital Engineering revolution, NASA and other space organizations are starting to embrace key concepts of model-based systems engineering and understand the value of moving from unstructured text documents to more formal knowledge representations that are amenable to automated data processing. In this line, this thesis leverages transformer models, a recent advance in natural language processing, to demonstrate automatic extraction of science relevance and traceability information from unstructured mission concept proposals. By doing so, this work helps pave the way for future applications of natural language processing to support other systems engineering practices within mission/program development such as automated parsing of design documentation. The proposed tool, called AstroNLP, is evaluated with a case study based on the Astrophysics Decadal Survey.

DEDICATION

I dedicate this work to my late father, as his guidance early in my life built me into the person I am today.

ACKNOWLEDGEMENTS

I would like to thank my committee chair, Dr. Selva, and my committee members, Dr. Dunbar and Dr. Huang, for their guidance and support throughout the course of this research as well as Dr. David Richardson out of NASA Goddard Space Flight Center for providing critical insight into NASA work practices and serving as my NASA-affiliated research collaborator.

Thanks also go to my friends for supporting me throughout this graduate experience and keeping the positivity high amidst this pandemic.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a thesis committee consisting of Professor Daniel Selva (advisor) and Professor Bonnie Dunbar of the Department of Aerospace Engineering and Professor Ruihong Huang of the Department of Computer Science and Engineering.

The training data discussed in Chapter 4 was provided, in part, by undergraduate student Kevin Zhang of the Department of Aerospace Engineering.

Funding Sources

This work was supported by a NASA Space Technology Graduate Research Opportunity Award (grant no: 80NSSC20K1226).

NOMENCLATURE

ADS	Astrophysics Data System
AI	Artificial Intelligence
API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
ML	Machine Learning
NASA	National Aeronautics and Space Administration
NER	Named-Entity Recognition
NLP	Natural Language Processing
NSSDCA	NASA Space Science Data Coordinated Archive
NSTGRO	NASA Space Technology Graduate Research Opportunity
PDF	Portable Document Format
POS	Parts of Speech
PRF	Precision, Recall, F1
P-STAF	Project-domain and Science Traceability Alignment Framework
STG	Science Traceability Graph
STM	Science Traceability Matrix
TRL	Technology Readiness Level

TABLE OF CONTENTS

	Page
ABSTRACT	ii
DEDICATION	iii
ACKNOWLEDGEMENTS	iv
CONTRIBUTORS AND FUNDING SOURCES.....	v
NOMENCLATURE.....	vi
TABLE OF CONTENTS	vii
LIST OF FIGURES.....	ix
LIST OF TABLES	xii
CHAPTER I INTRODUCTION	1
Overview	1
Relevance to NASA’s Technology Area Breakdown Structure (TABS)	2
CHAPTER II SCIENCE TRACEABILITY AND SCIENTIFIC RELEVANCE	3
Early Mission Concepts	3
Science Traceability	4
Science Traceability Matrix	5
Project-domain Science Traceability and Alignment Framework	9
Scientific (Programmatic) Relevance.....	10
Astrophysics Decadal Surveys	12
Overview	12
Science Panels and Questions	14
CHAPTER III NATURAL LANGUAGE PROCESSING.....	19
Overview	19
Ontologies	19
Semantic Strategies	20
Tokenization	21
Named-Entity Recognition.....	22

Relation Extraction.....	25
Transformer Architecture and BERT	26
Parts of Speech Tagging.....	29
Term Frequency and Topic Modeling.....	30
Performance Evaluation	31
Adapted PRF Metrics (MUC-5)	32
Research Question.....	33
CHAPTER IV NATURAL LANGUAGE PROCESSING FOR MISSION CONCEPT EVALUATION	35
Science Traceability Extraction	35
Science Traceability Graph Ontology	37
Document Text Extraction	41
Named-Entity Recognition Transformer.....	42
Relation Extraction Transformer.....	44
Tool Training.....	45
Graph Generation and Visualization	49
Science Relevance Assessment.....	50
Science Panels and Questions Knowledge Base	50
Noun-Chunking and Term Frequencies	51
Application to the Astrophysics Decadal Survey.....	52
Tool Graphical User Interface.....	52
System Performance.....	57
Graph Generation and Relevance Examples	59
Discussion	78
CHAPTER V CONCLUSIONS	80
REFERENCES	82
APPENDIX A SCIENCE PANEL AND SCIENCE QUESTION TOPIC/TERM LISTS	90

LIST OF FIGURES

	Page
Figure 1: The contents of a science traceability matrix as defined in [4] which provides the governing structure for all science traceability matrices.	6
Figure 2: Science traceability matrix of the Cosmic Dawn Intensity Mapper (CDIM) mission concept developed through NASA as a part of the 2020 Astrophysics Decadal Survey [5].	7
Figure 3: Science traceability matrix of the STROBE-X mission concept developed through NASA as a part of the 2020 Astrophysics Decadal Survey [5].	8
Figure 4: Toy example depicting the scientific impact of several distinct science themes based upon their perceived weight.	11
Figure 5: The task breakdown of the 2020 Astrophysics Decadal Survey as discussed in [9].....	13
Figure 6: Visualization of an embedding space (image source: towardsdatascience.com).	27
Figure 7: Scoring categories established in [51].	33
Figure 8: AstroNLP software architecture showing functions, data sources, process/data flows, and accompanying open-source libraries/tools.	36
Figure 9: Functional work-flow showing the processing pipeline of the AstroNLP tool.	37
Figure 10: Science traceability graph ontology governing the AstroNLP system as seen in [10]. This governing ontology serves as the guiding template for annotations.	40
Figure 11: UBIAI's graphical user interface when viewed from the document annotation portal [57].	46
Figure 12: Illustration of the input and output expectations for the AstroNLP tool.	50
Figure 13: AstroNLP's graphical user interface.	53
Figure 14: Region 1's document metrics across the repository of mission concepts. Here, the total number of documents, entities, tokens, and relations are provided for visual inspection to the user.	54

Figure 15: Region 2 of the AstroNLP tool showing specific science panel/question relevancies for the LUVOIR concept.	54
Figure 16: Region 3's science traceability graph's metrics for a specific mission concept (LUVOIR in this example). Notice that both the entity and relation lists can be viewed by selecting the appropriate tab.....	55
Figure 17: Region 4's relevance assessment panel. Here, the histogram plots can be viewed across two portfolios under comparison detailing the impacts these portfolios have on all/select science questions (science questions are listed via their ID number and printed after assessing the 'portfolio').....	56
Figure 18: Signature science cases for the LUVOIR concept as provided in [5].	61
Figure 19: STG extracted from the LUVOIR concept proposal using a higher-epoch (30) NER transformer.	63
Figure 20: LUVOIR's complete relevance distribution normalized across the decadal science questions.	64
Figure 21: LUVOIR's most relevant science panel as well as its top three most relevant science questions.....	65
Figure 22: LUVOIR's topic/term distributions for its most relevant science question. Notice the significant relevant representation of 'infrared' and 'planet' terms.	66
Figure 23: OST's science goals and science objectives as seen in [5].	67
Figure 24: OST's science traceability graph. A much larger version of this graph can be obtained if the baseline NER transformer is switched with a higher epoch model (i.e. one that went through all 30 cycles of the training data set).	68
Figure 25: OST's complete relevance distribution normalized across all decadal science questions.	69
Figure 26: OST's most relevant science panel and top three related science questions...	70
Figure 27: OST's topic/term distributions across its most relevant science question.	70
Figure 28: A portion of GEP's science theme and science action regions contained within its larger STG.	72
Figure 29: GEP's relevance distribution normalized across all decadal science questions.	73

Figure 30: GEP's most relevant science panel and top three most relevant science questions.	73
Figure 31: GEP's term/topic distribution across its most relevant science question.	74
Figure 32: Portfolio showing the 'loaded' concept portfolios ready for relevance analysis.	75
Figure 33: Each portfolio's relevance across all science questions (the upper is Portfolio 1 and the lower is Portfolio 2). This is, in essence, a concatenation of the individual relevance profiles of all mission concepts contained within either portfolio.	76
Figure 34: Topic/term distributions for science questions DQ-1 (under 'Panel on Galaxies').....	77
Figure 35: The comparison charts showing how portfolio 2's science impact compares with portfolio 1's scientific impact.	78
Figure 36: Panel and question portion of the knowledge based used as a reference guide for users of the AstroNLP tool so that relevance charts can be specifically pinpointed to specific scientific areas.	90
Figure 37: Term/topic map portion of knowledge base. Each row corresponds to a particular science question provided in the panel and question sheet of the excel document.	91

LIST OF TABLES

	Page
Table 1: The 2020 Astrophysics Decadal Survey's science panels, their science questions and their discovery areas as reported in [9].	15
Table 2: Various POS tags that an extracted token may be categorized as given spaCy's POS tagger model (https://spacy.io/usage/linguistic-features).	29
Table 3: List of all 10 entity types with associated descriptions and examples as seen in [10].	38
Table 4: Named entity recognition model parameters.	43
Table 5: Relation extraction model parameters.	44
Table 6: Training data size across all entity and relation types acquired over the course of 12 months.	47
Table 7: Profile analytics across all ten entity types.	48
Table 8: Baseline performance metrics for both transformer models based upon gold annotations.	58
Table 9: Full pipeline performance metrics for both transformer models. This scoring was carried out semi-automatically with relation extraction performance based off of NER model output as opposed to the gold annotations. Additionally, this scoring procedure only used ~25% of the testing data set (roughly 5% of the total training data set).	59
Table 10: Notable technical design features of the LUVOIR concept as seen in [5].	61
Table 11: Select technical details of the OST concept [5].	67
Table 12: Select design features of the GEP mission concept [5].	71

CHAPTER I

INTRODUCTION

Overview

Mission concept evaluation takes place across various domains in the aerospace enterprise (defense, planetary science, Earth science, and astrophysics to name a few). It is an important period within the systems engineering lifecycle of a program as it takes in a pool of alternative concepts aiming to achieve one or more programmatic objectives relevant to the wider program itself. By practice, it is up to the review panel(s) to determine what these programmatic areas are and determine/recommend which set of concept ideas should be considered for further development and implementation.

In this thesis, we aim to embed one subtopic of artificial intelligence into the realm of mission concept evaluation. Specifically, we will demonstrate the implications and benefits that natural language processing, an umbrella term that contains various semantic strategies and techniques ultimately aimed at analyzing and processing text, can bring to the area of mission concept evaluation.

Throughout this work, we will discuss AstroNLP, a tool developed to support the reviewer involved with assessing and issuing recommendations for mission concepts submitted to the Astrophysics Decadal Survey. Specifically, we will discuss the important implications that science traceability and scientific relevance have on the decadal process, discuss important tools and techniques contained within natural language processing, describe how such workflows were implemented in AstroNLP, and

demonstrate how they can be used in the decadal review process. Ultimately, as our use case, we will analyze three mission concepts submitted to the 2020 Astrophysics Decadal Survey and provide the outputs that AstroNLP generates when applied to these mission concept proposals.

Relevance to NASA's Technology Area Breakdown Structure (TABS)

This thesis is contained entirely within a NASA Space Technology Graduate Research Opportunity (NSTGRO). As such, the work discussed here is primarily motivated through NASA work practices and areas of growing technical interest at NASA. Specifically, this thesis serves to address portions of NASA Technology Area 11.4.3 Semantic Technologies and NASA Technology Area 11.4.4 Collaborative Science and Engineering [1]. Thematically, the work discussed in this thesis addresses these two areas through applications of semantic technology (i.e. natural language processing) to provide intelligent data understanding within the context of mission concept evaluation and support the already collaborative engineering and scientific effort that is the decadal process. Furthermore, and throughout the duration of this thesis, much of the motivation driving the goals and direction of this work was formulated by these NASA TABS areas and through conversations with NASA decadal experts and research collaborators. However, it should also be noted that the implications of this work hold relevance to other similar activities across various engineering-related enterprises and as such, can be adapted to said domains.

CHAPTER II

SCIENCE TRACEABILITY AND SCIENTIFIC RELEVANCE

Early Mission Concepts

Mission concepts, or what can be more formally considered as pre-Phase A mission concepts, hold critical mission and design information that is necessary to communicate what are the goals and intended impacts of said mission. As stated in the NASA Systems Engineering Handbook, pre-Phase A mission concepts are an essential part of concept studies aimed at producing a variety of different ideas and alternatives through which a program can be built from [2]. These concepts can range from proposed projects and missions aimed at serving one or more programmatic goals and objectives, or can be even more specific targeting high risk and/or low TRL technologies that could serve programmatic objectives/goals in whichever way the proposer identifies [2]. Once these various concepts, technologies, and proposed activities are reported upon and submitted to a program for review, it is here where a critical step in the systems engineering lifecycle begins: program formulation [2].

In program formulation, review panels are formed and implemented so as to sift through the various proposals submitted to the program and down-select a promising program of concepts from the original submission pool. Here, several pieces of information are important when considering a concept for a program. These can include (but are not limited to):

- Proposed mission/concept/technology development cost

- Proposed development/implementation schedule
- Technical feasibility and merit
- **Relevance of mission concept to programmatic goals and objectives**
- **Traceability of requirements and engineering decisions from programmatic goals and objectives**
- Clarity and completeness of proposal's content

The specific details and figures of merit pertaining to each of these evaluation criteria are dependent upon the specified program. For example, military-related programs may focus on functional capability of new technologies that provide further support to existing U.S. defense capabilities and/or are of direct benefit to the warfighter [3]. Conversely, in the case of science based-missions, considerations of relevance to the scientific program and traceability of engineering decisions/requirements from scientific goals and objects are undoubtedly important (in addition to the other criterion listed above).

For the work covered in this thesis, we wish to focus on the two bolded points listed above as our areas of focus. As such, we will expand on the importance of science traceability in the following section.

Science Traceability

When considering traceability, particularly in requirements, it's important that such definitions are constructed coherently, concisely, and mapped effectively so that reviewers, stakeholders, and systems engineers alike can adequately adjudicate the rationale behind why such requirements exist and where they come from (as well as

illustrate, to a degree, how said requirements meet the programmatic goals and objectives) [2]. In science-based missions, this is commonly referred to as science traceability and is the basis behind how a mission illustrates its potential in meeting one or more scientific goals and objectives. One method of establishing science traceability, and is requirement for all NASA mission concepts, is an artifact called the science traceability matrix [4].

Science Traceability Matrix

The science traceability matrix, first reported in [4], is a useful way in formulating a mission's science traceability structure in a tabular format. Here, the mapping of traceability is illustrated in a left-to-right and right-to-left flow (bidirectional traceability) with higher level definitions (e.g. science goals and objectives) located on the left hand side of the table whilst lower level definitions specific to the mission concept (e.g. instrument and mission requirements) are located on the right hand side of the table. Also accompanying these elements are measurement objectives, measurements requirements, and various items which can be categorized as data products located more towards the center of any given matrix. The following figure illustrates this general traceability flow:

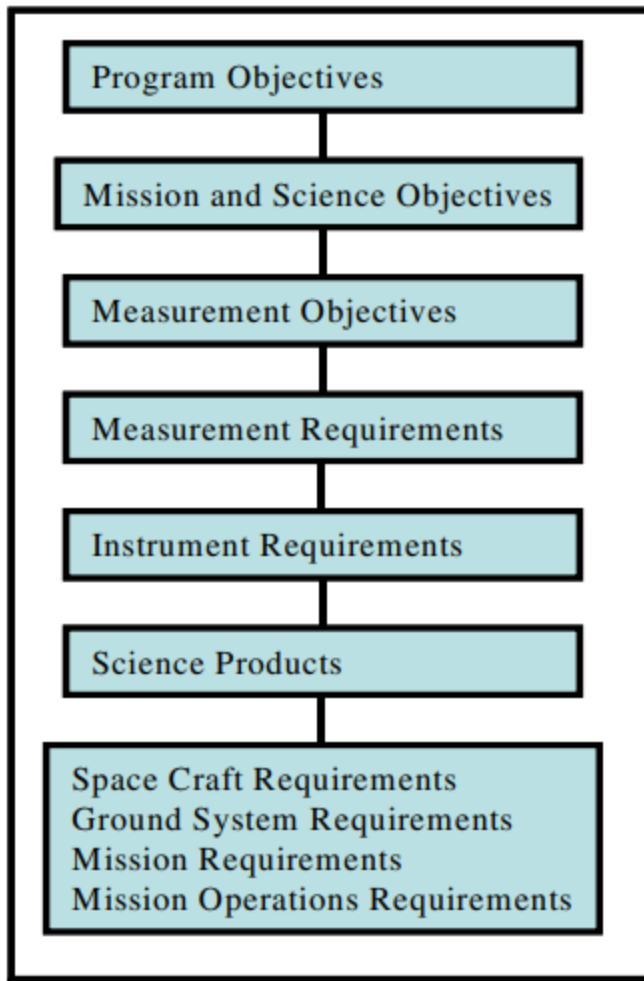


Figure 1: The contents of a science traceability matrix as defined in [4] which provides the governing structure for all science traceability matrices.

As the author states in [4], it is important for all science traceability matrices to cover what are the baseline scientific goals and objectives given by the mission concept (or as the author states, an ‘Announcement of Opportunity’) that are relevant to a program. From here, it is important for said concepts to then trace this high-level information to measurement requirements, of which is then to be traced to more specific instrument/mission/spacecraft requirements of which will be implemented to satisfy such

measurement requirements. To illustrate examples of science traceability matrices, consider the following two figures:




NASA Science Goals	CDIM Science Goals	CDIM Science Objectives	Science Requirements			Instrument Requirements			Mission Requirements	
			Physical Parameters	Observables	Measurement Requirement	Instrument Parameter	Science Requirement	Capability	Driver	Parameter
<p>Explore the origin and evolution of the galaxies, stars and planets that make up our universe [NASA Science Plan]</p> <p>How does the Universe work? How did we get here? [NASA 2014 Science Mission Directorate Strategy Document]</p>	<p>Trace the stellar mass buildup, dust production history, and metal enrichment history during cosmic reionization.</p>	<p>Determine if the rate of growth of metals and dust corresponds to the growth of stellar mass at $5 < z < 8$.</p> 	<p>Metallicity of galaxies via the oxygen abundance, stellar mass, and dust attenuation (extinction rate, dust density)</p>	<p>[OIII], [OII], [NII]Hα, Hα/Hβ @ $5 < z < 8$</p>	<p>(i) Wavelength coverage to detect Hα out to z of 10. (ii) Spectral resolving power to resolve [NII] and Hα. (iii) Sensitivity to detect galaxies $< 10^9 M_{\text{sun}}$ in a deep survey.</p>	Wavelength range	$2.2 \leq \lambda \leq 6.0 \mu\text{m}$	$0.75 \leq \lambda \leq 7.5 \mu\text{m}$	<p>Deep, medium and wide surveys each with $\geq 90\%$ voxel completeness for internal reliability. Spatial resolution: Effective PSF FWHM $\leq 2''$ at $1 \mu\text{m}$ (from science requirements). Stable cooling to $< 35 \text{ K}$ to control $> 5 \mu\text{m}$ array dark current.</p>	
						Spatial resolution (pixel scale)	$\Theta_{\text{pix}} = 1''\text{--}2''$	$\Theta_{\text{pix}} = 1''$		
						Spectral resolving power	$\lambda/\Delta\lambda \geq 300$	$\lambda/\Delta\lambda \geq 300$		
<p>Establish the role of active galactic nuclei (AGN) in cosmic reionization.</p>	<p>Determine the fractional contribution of super-massive black hole/AGNs to reionization photon budget.</p> 	<p>Unbiased UV photon spectral density, black-hole masses via line widths of optical lines.</p>	<p>Rest-frame UV continuum @ $z = 5\text{--}8$. [MgII] and other metal lines.</p>	<p>(i) Sensitivity to detect faint quasars in a wide survey. (ii) Spectral resolving power to detect equivalent width of broad metal lines.</p>	Wavelength range	$2.9 \leq \lambda \leq 6.0 \mu\text{m}$	Same as above	<p>Deep survey: 15 deg^2, imbedded in the Wide survey. Medium survey: 30 deg^2, to overlap with 21-cm fields from HERA and SKA1-Low. Wide survey: 300 deg^2, driven by number of AGN detections. Read, reduce, and telescope spectral imaging data.</p>		
					Spatial resolution (PSF, FWHM)	$\Theta_{\text{FWHM}} = 2''$ at K band	$\Theta_{\text{FWHM}} < 2''$ at K band			
					Spectral resolving power	$\lambda/\Delta\lambda \geq 300$	Same as above			
<p>Establish the progression and topology of reionization from cosmic dawn at $z = 10$ to the end of reionization at $z < 6$.</p>	<p>Determine the progress of reionization by measuring the ionization fraction in at least 10 redshift bins at $5 < z < 10$, with accuracy better than 10%.</p> 	<p>Lyα luminosity function, escape fraction, and the spatial distribution.</p>	<p>Lyα</p>	<p>(i) Wavelength coverage to detect Lyα out to z of 10. (ii) Sensitivity to detect faint galaxies.</p>	Wavelength range	$0.75 \leq \lambda_{\text{Ly}\alpha} \leq 0.98 \mu\text{m}$	$0.75 \leq \lambda \leq 7.5 \mu\text{m}$	<p>Wide survey: 300 deg^2, driven by number of AGN detections. Read, reduce, and telescope spectral imaging data.</p>		
					Spectral resolving power	$\lambda/\Delta\lambda \geq 100$	Same as above			
					Spectral line flux sensitivity (3.5 σ ; deep survey)	$2.9 \times 10^{-17} \text{ erg s}^{-1} \text{ cm}^{-2}$ at $0.85 \mu\text{m}$	$2.0 \times 10^{-17} \text{ erg s}^{-1} \text{ cm}^{-2}$ at $0.85 \mu\text{m}$			
<p>Reionization history of the universe.</p>	<p>Lyα and Hα.</p>	<p>Lyα and Hα.</p>	<p>(i) Ability to perform cross-correlations, including Lyα and Hα, and 21-cm radio measurements.</p>	Wavelength range	$0.75 \leq \lambda_{\text{Ly}\alpha} \leq 1.4 \mu\text{m}$ $3.9 \leq \lambda_{\text{H}\alpha} \leq 7.2 \mu\text{m}$	$0.75 \leq \lambda \leq 7.5 \mu\text{m}$	<p>Wide survey: 300 deg^2, driven by number of AGN detections. Read, reduce, and telescope spectral imaging data.</p>			
				Spectral resolving power	$\lambda/\Delta\lambda \geq 100$	Same as above				
				Surface brightness sensitivity (1 σ ; medium survey)	$1.3 \times 10^{-18} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ Hz}^{-1} \text{ sr}^{-1}$ at $1.1 \mu\text{m}$	$1.5 \times 10^{-18} \text{ erg s}^{-1} \text{ cm}^{-2} \text{ Hz}^{-1} \text{ sr}^{-1}$ at $1.1 \mu\text{m}$				

Figure 2: Science traceability matrix of the Cosmic Dawn Intensity Mapper (CDIM) mission concept developed through NASA as a part of the 2020 Astrophysics Decadal Survey [5].

Science Goals	Science Objectives	Scientific Measurements	Driving Requirements
1. Measure the spin distribution of accreting black holes	1.1 Measure the spin distribution of accreting black holes	Thermal continuum	Energy range: 0.2–30 keV
		Reflection & X-ray reverberation	Energy resolution: 200 eV
	1.2 Measure BH spin for 20 AGN to <10%	High frequency QPO	Time resolution: 100 microsec
		Transient outbursts	Effective Area: 20,000 cm ² Observe bright sources with full energy and time resolution
2. Understand the equation of state of dense matter	2.1 Measure the mass and radius to within 5-10% for ~20 pulsars to map the EOS and probe potential phase transitions	Reflection & X-ray reverberation	Wide-field monitoring: 75% of sky, 5 mcrab (1 day) sensitivity, 1 keV energy resolution, 2 arcmin position accuracy ToO response (< 24 hours)
		Jetted TDE detection	Energy range: 1–30 keV Energy resolution: 200 eV
	2.2 Search for the fastest spinning pulsars	Pulse profile modeling for rotation powered pulsars, accretion powered pulsars, and thermonuclear burst oscillation sources	Effective Area: 20,000 cm ²
3. Explore the properties of the precursors and electromagnetic counterparts of gravitational wave sources	3.1 Enable detection of 5–10 short gamma-ray bursts per year	Search for spin frequencies up to 2 kHz	Effective area: 16,300 cm ² @ 1 keV/38,200 cm ² @ 6 keV; Time resolution: 80 microsec Energy resolution: 85-175 eV FWHM (0.2-10 keV) TOO response time: hours
	3.2 Search for signatures of merging supermassive BH	Detect and localize w/ immediate trigger or ground searches	Time resolution: 50 microsec
			Wide-field monitor as above with 1ms time resolution All wide-field monitor data downlinked to ground

Figure 3: Science traceability matrix of the STROBE-X mission concept developed through NASA as a part of the 2020 Astrophysics Decadal Survey [5].

As will be discussed later and can be seen visually through these examples, no two STMs are the same which is obviously true given that STMs usually cover one specific mission concept each. However, the STM, whilst a standard requirement for NASA proposals, is not standardized in-of-itself. Henceforth, the degree of variability between two STMs is immense as two STMs could contain wide varieties in data representation, tabular structure, and completeness to name a few (this is true even for STMs within the same program as illustrated above).

This implication of variability brings complications from the perspective of a concept reviewer working to fill their scientific program (e.g. a panel reviewer who adjudicates mission concepts for various decadal surveys such as the Earth Science and

Astrophysics decadal surveys). Further still, one must consider whether or not this artifact is available to the reviewer especially when considering pre-Phase A mission concepts under review by a program. In those cases, of which will be discussed later, the degree of variability is further amplified as such traceability information, and structure, is buried within the text of a concept proposal. This undoubtedly has implications for reviewer quality and raises the chance that mission critical information could be missed on the behalf of the reviewer [6].

As consequence of this, it is worthy to mention prior work that has catapulted off the foundation that [4] provided in establishing the STM. One such work develops the concept of the Project-domain Science Traceability and Alignment Framework [7].

Project-domain Science Traceability and Alignment Framework

The authors in [7] identified that the STM does not necessarily provide a complete mapping of science traceability as required for mission concept evaluation. As such, they provided an extension of the science traceability matrix through the implementation of “common definitions and valid relations to structure the communication across the project” [7]. Specifically, the Project-domain Science Traceability and Alignment Framework (P-STAF) borrows the governing structure of the STM and adds other fields fundamental to the framework of P-STAF.

More notably, from the perspective of the work contained in this thesis, the authors in [7] showed a graph-based representation of their P-STAF artifact, through what was referred to as a STAF science information network. This network detailed the mappings of STM and P-STAF elements through various nodes and edges as a part of

their case study on applying the P-STAF framework towards a planned Europa mission [7].

This reformulation of the STM into P-STAF shows the benefit of establishing further semantic structure to such an important element of a mission concept (i.e. its science traceability). Additional work regarding P-STAF looked at an analysis of payload architectures, covered in [8], which provides a further representation of the P-STAF taxonomy. As will be discussed more thoroughly in following chapters, these follow-ons to the STM represent the value that refined taxonomies, or more generally ontologies, bring when formulating science traceability.

The next item of discussion pertinent to this work is in regard to assessing scientific relevance of a mission concept, which cannot be fully explained under the topic of science traceability.

Scientific (Programmatic) Relevance

When building a scientific program of science-based missions, the reviewer must consider not only if the engineering decisions and projected data products are mapped appropriately to the proposal's defined science goals and objectives, but whether or not those science goals/objectives are relevant to the science program altogether. What must also be considered is whether or not a selected set of candidate concepts cover the scientific themes or capabilities in way that is considered desirable by the review panel. In essence, relevance can be considered as way of providing coverage across all programmatic goals or, on the other hand, targeting specific areas of high interest more heavily and areas with low interest less heavily.

A more formal way to address the matter of relevance can be issued through the lens of probability distribution. Take for example the two following distributions where the height of each bar in the chart represents the degree of interest or scientific impact a certain scientific theme (i.e. science goal/objective) has to the program:

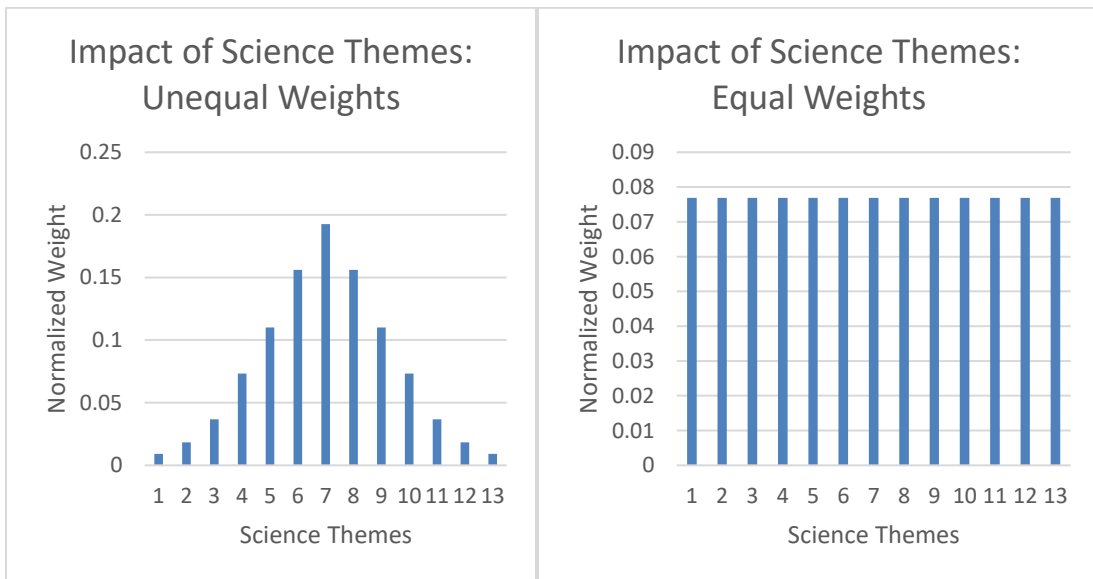


Figure 4: Toy example depicting the scientific impact of several distinct science themes based upon their perceived weight.

As the above figure illustrates, a scientific program may institute a bias, either deliberately or unintentionally, when adjudicating the importance of science themes. As such, this directly affects the relevance of a portfolio of candidate missions as a portfolio may be valued as significantly relevant in one distribution but less so in another.

Through the considerations of science traceability and scientific relevance, it was determined early in the project that a ripe case study for this work would exist in the

realm of decadal surveys. Specifically, the Astrophysics Decadal Survey conducted by the National Academies of Science, Engineering, and Medicine.

Astrophysics Decadal Surveys

Overview

The decadal surveys hosted by the National Academies of Science, Engineering, and Medicine, as stated above, aim to provide recommendations for scientific programs for the following decade in a way that best meets the goals and objectives of the related community. For the Astrophysics Decadal Survey in particular, of which has recently released its 2020 report and can be found in [9], the survey focused on establishing thematic areas in astrophysics and generated specific recommendations across those thematic areas through a series of reviews across several panels. The following figure summarizes the statement of task that governed the survey:

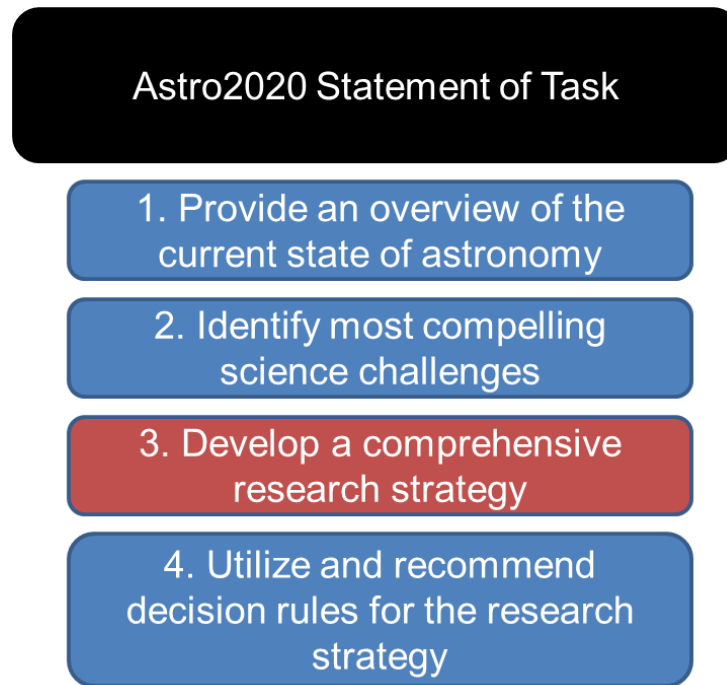


Figure 5: The task breakdown of the 2020 Astrophysics Decadal Survey as discussed in [9].

Considering the specific tasks of extracting science traceability and determining scientific relevance of mission concepts (so as to create an impactful portfolio), the work in this thesis is primarily associated with task area three shown in the above figure. As stated in [9], task area three aims to “develop a comprehensive research strategy to advance the frontiers of astronomy and astrophysics for the period 2022-2032”. Further, this task area will consider the science case for each proposed activity (‘activity’ including large, medium, or small ground or space-based research programs) and generate recommendations on which activities should be considered for the upcoming decade [9].

For these reasons, and as this title of this thesis establishes, the aim of this thesis work was to develop a tool that can demonstrate the potential benefit that natural language processing can bring to a mission concept evaluation effort. We have recently reported on a portion of this work focusing on extracting science traceability graphs using natural language processing in [10]. There has also been prior work discussed in [11] that examined topic trends across the decadal survey's history in an effort to improve the science prioritization process. Specifically, that work used Latent Dirichlet Allocation to (LDA) capture astrophysics-based topic frequencies and temporal trends by analyzing prior submissions to the decadal [11]. Beyond these two reported ventures however, little literature exists on bringing natural language processing to the astrophysics decadal survey *process* and, as such, is a gap we aim to address in this thesis.

Science Panels and Questions

One final discussion regarding the decadal survey process and of which is crucial to the task of determining scientific relevance is in regards to the decadal's science panels and science questions. In the most recently released decadal survey for astrophysics [9], the decadal established six science panels targeting various scientific themes related to astronomy and astrophysics. Those six panels were:

1. Panel on Compact Objects and Energetic Phenomena
2. Panel on Cosmology
3. Panel on Galaxies
4. Panel on Exoplanets, Astrobiology, and the Solar System

5. Panel on the Interstellar Medium and Star and Planet Formation
6. Panel on Stars, the Sun, and Stellar Populations

Contained within the decadal survey are six appendices associated with each of these science panels of which contain the reports on the findings of each panel [9]. In each of these appendices, four unique science questions are posed per panel (with a varying amount of subsidiary science questions/topics for each of these four questions) in addition to one area of discovery, and a breakdown of capabilities and future needs necessary to serve these science questions [9]. The following table outlines these science questions and discovery areas per panel as seen in the decadal survey:

Table 1: The 2020 Astrophysics Decadal Survey's science panels, their science questions and their discovery areas as reported in [9].

Panel on Compact Objects and Energetic Phenomena	
B-Q1	What are the mass and spin distributions of neutron stars and stellar-mass black holes?
B-Q2	What powers the diversity of explosive phenomena across the electromagnetic spectrum?
B-Q3	Why do some compact objects eject material in nearly light-speed jets, and what is that material made of?
B-Q4	What seeds supermassive black holes and how do they grow?
B-DA	Transforming our view of the universe by combining new information from light, particles, and gravitational waves

Panel on Cosmology	
C-Q1	What set the Hot Big Bang in motion?
C-Q2	What are the properties of dark matter and the dark sector?
C-Q3	What physics drives the cosmic expansion and large-scale evolution of the universe?
C-Q4	How will measurements of gravitational waves reshape our cosmological view?
C-DA	The Dark Ages as a cosmological probe
Panel on Galaxies	
D-Q1	How did the intergalactic medium and the first sources of radiation evolve from cosmic dawn through the epoch of reionization?
D-Q2	How do gas, metals, and dust flow into, through, and out of galaxies?
D-Q3	How do supermassive black holes form and how is their growth coupled to the evolution of their host galaxies?
D-Q4	How do the histories of galaxies and their dark matter halos shape their observable properties?
D-DA	Mapping the circumgalactic medium and intergalactic medium in emission
Panel on Exoplanets, Astrobiology, and the Solar System	
E-Q1	What is the range of planetary system architectures and is the configuration of the solar system common?

E-Q2	What are the properties of individual planets and which processes lead to planetary diversity?
E-Q3	How do habitable environments arise and evolve within the context of their planetary systems?
E-Q4	How can signs of life be identified and interpreted in the context of their planetary environments?
E-DA	The search for life on exoplanets
Panel on the Interstellar Medium and Star and Planet Formation	
F-Q1	How do star-forming structures arise from, and interact with, the diffuse interstellar medium?
F-Q2	What regulates the structure and motions within molecular clouds?
F-Q3	How does gas flow from parsec scales down to protostars and their disks?
F-Q4	Is planet formation fast or slow?
F-DA	Detecting and characterizing forming planets
Panel on Stars, the Sun, and Stellar Populations	
G-Q1	What are the most extreme stars and stellar populations?
G-Q2	How does multiplicity affect the way a star lives and dies?
G-Q3	What would stars look like if we could view them like we do the Sun?
G-Q4	How do the Sun and other stars create space weather?
D-DA	“Industrial-scale” spectroscopy

These science panels and questions, in addition the subtopics related to these questions not shown above, form the foundation of determining the relevance of a research program to the decadal. The utility of these guiding topics will be discussed more thoroughly in Chapter 4.

CHAPTER III

NATURAL LANGUAGE PROCESSING

Overview

Natural language processing, which falls under the broader term of semantic technology, aims to provide tools and techniques capable of ingesting text in a variety of formats and producing usable data for further computation. As stated in [12], the rise of the world wide web has given way to a plethora of textual sources which are ripe for applications of natural language processing. Further still, applications of various techniques within natural language processing/semantic technology are on the rise in various engineering-related works such as requirements engineering [13-17]. NASA also outlines semantic technology as one of its low TRL technology areas through the 2020 Technology Taxonomy [18]. In this chapter, we will discuss various strategies of natural language processing as it relates to this thesis. First, however, we introduce the important topic of ontologies as it is the most fundamental element regarding our work.

Ontologies

By definition, ontologies are a formal representation of a domain containing elements of various types (e.g. classes, attributes, and relations) to describe said domain. As stated in [19], ontologies exist in a branch of metaphysics concerned with “science of being as being”. In modern practice, particularly in the space domain, ontologies have been used to describe specific sub-domains such as space objects [20], space debris [21], satellites databases [22], space systems [23], as well as for other less space-related

domains such as systems engineering [24], intelligence, surveillance, and reconnaissance [25-27] and requirements engineering [28-31].

As an example of an ontology, one can consider the Missions and Means Framework (MMF) Ontology defined in [26]. The main concepts of the MMF ontology are defined and categorized in a directed graph with nodes and edges. Another layer of this ontology can be visualized by grouping certain nodes and edges to sub-regions within graph which is reported in a following work [27].

When considering ontologies in the realm of this thesis, it was determined early that ontologies are widely useful in providing a formal description of the domain of application (in this case, astrophysics-based space missions). Further, creating a way of embedding natural language processing to support mission concept evaluation particularly in extracting science traceability and evaluating scientific relevance, we determined that the formation of a domain-specific ontology was necessary. We will discuss the details of our developed ontology in the following chapter as well as make mention of reference ontologies/taxonomies used to create said ontology.

With ontologies providing the foundation of the semantic technology employed in this thesis, the next section will discuss specific strategies and techniques commonly used in natural language processing that were henceforth employed in our processing pipeline.

Semantic Strategies

All of the semantic processing done in this domain of application will act upon unstructured textual sources (as described in the previous chapter and is of direct

consequence of our domain of study). As such, there exists a need to define the elements necessary to form a processing pipeline so as to transform those original chunks of text into useful data for the reviewer. Some of the first steps taken on these textual samples hardly warrant much background discussion here and are better discussed in the following chapter (i.e. document and sentence segmentation). This brings us to the first major processing milestone, tokenization.

Tokenization

For simplicity, a ‘token’ is a collection of characters that, through a human’s eye, form a word, phrase, or simply punctuation. To a computer however, these tokens are fundamental units used for processing that are then manipulated in such a way that downstream tasks can provide further information about the original textual sample. Consider the following sentence taken from [16] detailing the input and output of a sentence to its tokenized form:

- INPUT
 - “These prerequisites are known as (computer) system requirements and are often used as a guideline as opposed to an absolute rule.”
- OUTPUT
 - [‘These’, ‘prerequisites’, ‘are’, ‘known’, ‘as’, ‘(’, ‘computer’, ‘)’’, ‘system’, ‘requirements’, ‘and’, ‘are’, ‘often’, ‘used’, ‘as’, ‘a’, ‘guideline’, ‘as’, ‘opposed’, ‘to’, ‘an’, ‘absolute’, ‘rule’, ‘.’]

In the above example, there is a clear subdivision of the original sentence into its constituent words and punctuation. The authors in [16] used the python-based Natural Language Toolkit (NLTK) [32] for this tokenization task but there are other open-source options available including spaCy [33] which is used in our work. With these words now in a tokenized form, it is important to ‘clean’ these tokens, or more formally lemmatize them, so that the extra noise intrinsic to the language (e.g. plurals, capitalization, suffixes) are removed. For example, the lemmatized form of the token ‘opposed’ would be ‘oppose’, removing the ‘ed’ suffix by replacing it with its root form.

As was mentioned, there are several open-source natural language processing libraries available that can perform one or more natural language processing tasks. In many cases, these libraries utilize different types of models (e.g. transformers, neural networks, rules-based methods, and embeddings to name a few) for these tasks. In such cases, some models are better utilized in certain domains rather than others, particularly when considering pre-trained statistical models like neural networks and transformers. Therefore, it is important to recognize that although the above tokenization example is provided in the way it was described in [16], a different tokenization model may segment the sentence differently. Due to the implications of this, we shift the discussion now to named-entity recognition where the idea of using specially adapted models to specific tasks can better be explained.

Named-Entity Recognition

In named-entity recognition, a model aims to extract certain ‘entities’ from text and assign them a classifying label. There are several reported works detailing

applications of named-entity recognition in practice [34-41] as well as surveys covering several methods in named-entity recognition [42, 43].

To illustrate an example of named-entity recognition in practice, consider the following sentence:

- “The *Earth* {entity_type: *PLANET*} is a celestial body in orbit around the Sun.”

In the above sentence, the entity “Earth” is extracted and classified as a ‘PLANET’ entity type through our fictional named-entity recognition system. It is important to behold the fact that this supposed system used some form of strategy to not only classify the entity with a label (‘PLANET’) but also determine which set of tokens within the text were indeed an entity. Further still, it can be argued that other entities exist within the text (e.g. “Sun” could be classified as a ‘STAR’). However, a named-entity recognition system in this context was likely built to extract entities of various predefined entity types (assuming that the above output is 100% accurate to the model) meaning that other potential entity types recognized by a reader are ignored by the machine. This reiterates the importance of establishing a governing ontology which classifies these entity types so that further downstream tasks can be performed more effectively in that given domain.

In recent years, named-entity recognition has been largely employed through the use of neural networks [43] and transformer models. As stated in [43], a survey of named-entity recognition practices conducted in 2018, neural networks were shown to outperform more classical feature-engineered models. However, this survey did not

cover the more recent advance of transformer models discussed in [44] which have become the state-of-the-art for many named-entity recognition tasks.

In the case of statistical methods, particularly that of neural networks and transformers, named-entity recognition models typically employ some degree of training (or fine-tuning if pretrained models are available) so that the model is well adapted to its domain of application. There exist public datasets available for training named-entity recognition models, some of which used the ‘PERSON’, ‘LOCATION’, and ‘ORGANIZATION’ entity labels for training, as was the case in [34] (which used the Computational Natural Language Learning (CoNLL) 2003 and Open Knowledge Extraction (OKE) 2016 benchmark datasets). In several cases however, the named-entity recognition task for a certain problem may be unable to use these public datasets as the domain corpora and/or entity types are not useful for the specific domain of study. In these cases, it is necessary to determine what available corpora, and/or training datasets, are available so as to develop a custom named-entity recognition system that is best-suited for your task environment. For example, in [41], the authors were tackling the issue of extracting the Hubble Constant from a variety of related scientific texts. Henceforth, their custom-trained named-entity recognition system had to train off a custom-built training dataset consisting of 1,394 positive/negative training examples with 154 examples allocated towards an evaluation (testing) set [41]. It should be noted that the complexity of a named-entity recognition system, both in terms of number of entity types and degree of variability in entity types, has enormous implications on both

the performance of a model and the ‘necessary’ size of a training dataset needed to reach an optimal performance envelope (as discussed in [45]).

Finally, when considering name-entity recognition for a wider natural language processing task like information extraction, a useful follow-on technique comes in the form of relation extraction.

Relation Extraction

Relation extraction involves answering the question of whether a relationship exists between two entities and what type of relation that is. As stated in [46], a relation “denotes a well-defined relationship between two or more named entities”. For example, consider again the following sentence mentioned in the previous section but with an added entity type:

- “The *Earth* {entity_type: *PLANET*} is a celestial body in orbit around the *Sun* {entity_type: *STAR*}.”

In this sentence, the appearance of two entities brings forward a question of whether or not they are related and through what type of relation. As such, a relation extraction system could predict that the relation between these two entities does indeed exist and is of the type ‘ORBITS’. Thus, this completes the tuple containing two entities and a relation, i.e. “Earth”→ORBITS→“Sun”. Of course, not all entities have to be related to each other and any predictions should follow a predefined governing relation and entity type list (e.g. an ontology).

As is the case with named-entity recognition, there are several functional methods that can be employed for the task of relation extraction from supervised to

unsupervised [46]. Feature-based methods, kernels, bootstrapping, neural networks, and transformers are some of the various methods used/proposed for relation extraction as reported in various works and reviews [35, 36, 46, 47]. In a recent publication on this work, we specifically employed a transformer model to handle the task of relation extraction [10].

Given the rise of transformers models in the domain of natural language processing, and the specific presence of bi-directional encoder representations from transformers (BERT) in this work, it is warranted to provide discussion behind the inner-workings of this powerful method.

Transformer Architecture and BERT

The general transformer architecture follows an encoder-decoder processing pipeline and was initially conceptualized to support translation efforts as a replacement to slower long short term memory (LSTM) networks [48]. In the context of language translation tasks, the transformer architecture aims to take the input language sample and provide its translation as an output [48]. The role of the encoder is to transform the word embeddings (which are essentially vectors unique to each word in the vocabulary) into attention vectors that can be compared with the similar attention vectors in the other language [48]. Essentially, the encoder learns the context of the input language while the lower half of the decoder (in reference to the above diagram) does the same for the target language. With these attention vectors for both languages, they can be correlated to generate a word-by-word prediction of the translation in the target language, outputted as vector probabilities, which is then transformed into a more interpretable output for the

user [48]. One important element of this architecture is the use of word embeddings which, as just previously stated, is essentially a mapping of the entire domain vocabulary onto an embedding space. The following figure provides a visualization of what a low-dimensional embedding space can look like:

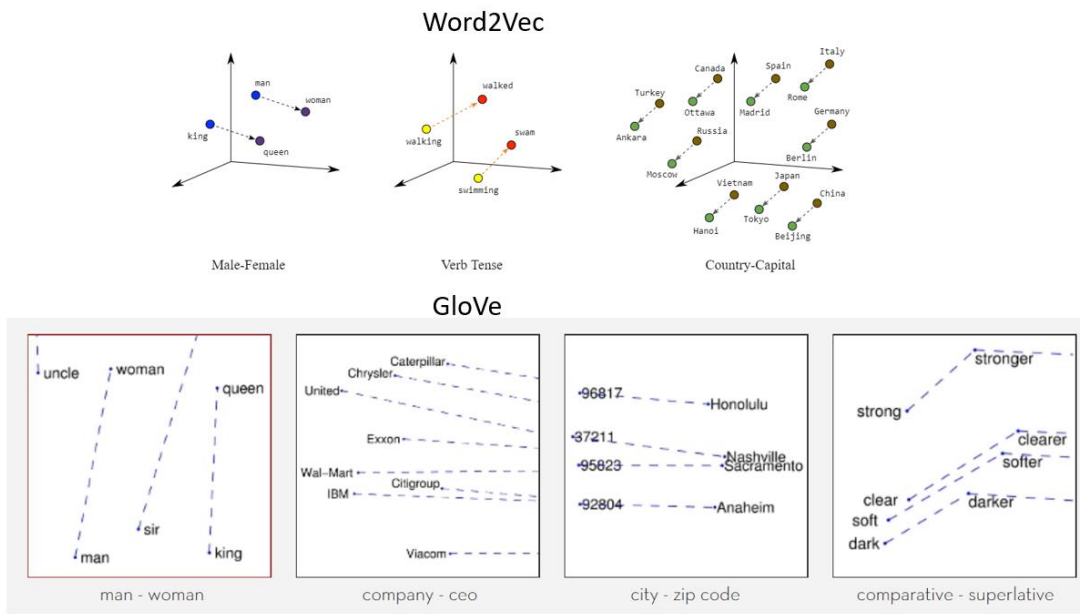


Figure 6: Visualization of an embedding space (image source: towardsdatascience.com).

Embeddings themselves can be useful for a variety of natural language processing tasks like entity disambiguation and relation extraction. For example, [49] used a noun-phrase embedding model to improve feature extraction in the engineering design domain.

To expand on the original transformer architecture, researchers at Google developed BERT, which utilizes the encoder block [44]. Here, the capabilities of the transformer, or rather the encoder block, can be expanded to support a wider variety of natural language processing tasks such as question and answering, text classification, and named entity recognition (to name a few).

The outputs of the encoding blocks can be reworked from the original architecture to serve other natural language processing tasks. In question and answering for example, the BERT model is pre-trained (unsupervised) on a corpus to ‘learn’ the context of the text vocabulary and structure [44]. This is done through Mask Language Modeling (MLM) which randomly masks certain words in a sentence [44]. The model then predicts which word lies behind mask [44]. The other element in pre-training, also unsupervised, is Next Sentence Prediction (NSP) which further trains the model by randomly gathering two sentences and training the model to predict which one comes after [44]. In all, the original BERT architecture encompasses 110 million (BERT base) or 340 million (BERT large) model parameters to train on and fine-tune [44].

Fine-tuning a pre-trained BERT model is task specific but rather straightforward [44]. Essentially, the model uses the NSP as a binary analog to a specific task input and the output provides start and ending spans (i.e. locations in the text) that indicate the model’s prediction [44]. In all, this architecture provides a very suitable tool for classification tasks like named entity recognition and relation extraction. Several works since the original publication of BERT have utilized this architecture to develop other

pre-trained models [36, 39] or have fine-tuned BERT to solve specific natural language processing tasks [10, 37].

Parts of Speech Tagging

Parts of Speech (POS) tagging in natural language processing is, in essence, the technique of assigning grammatical labels to extracted tokens. Similar to named entity recognition, there exist several models capable of performing automatic POS tagging on textual samples. The following are various POS tags, and examples, available through spaCy’s open-source POS tagger:

Table 2: Various POS tags that an extracted token may be categorized as given spaCy's POS tagger model (<https://spacy.io/usage/linguistic-features>).

POS Tag	Description	Examples
ADJ	Adjective	big, old, green
ADP	Adposition	of, to, from
ADV	Adverb	very, where, there
AUX	Auxiliary	is, has, will
CONJ	Conjunction	and, but, or
CCONJ	Coordinating Conjunction	and, but, or
DET	Determiner	a, an, the
INTJ	Interjection	psst, ouch, hello
NOUN	Noun	person, tree, air
NUM	Numeral	1, four, MMXIV
PART	Particle	‘s, ‘t
PRON	Pronoun	I, she, they
PROPN	Proper Noun	Ben, Sue, NATO

PUNCT	Punctuation	‘, (, ?
SCONJ	Subordinating Conjunction	if, while, that
SYM	Symbol	&, %, ©
VERB	Verb	run, eat, ate
X	Other	asdfsfs
SPACE	Space	

Term Frequency and Topic Modeling

One final task related to natural language processing and is arguably one of the simplest tasks to employ, is that of term frequency. Simply speaking, term frequency aims to analyze the occurrence of words throughout a corpus, or similarly determine the occurrence of specific words (or topics) in a corpus if said terms/topics are known. This, when used in methods such as term frequency/inverse document frequency (TF/IDF), Latent Dirichlet Allocation (LDA), or Latent Semantic Analysis (LSA), is useful to understand the ‘heatmap’ of words/topics within a document and generate topic models.

Several prior works have used different forms of term frequency analysis and topic modeling to solve various tasks. For example, the authors in [50] used both LDA and LSA to support the categorization of patents. LDA was also used in [11] to look at topic frequency and temporal trends across astrophysics decadal surveys (as mentioned in the previous chapter). However, while LDA can run unsupervised, the outputs of the model are dependent on the source text and cannot be defined a-priori. This makes it difficult for certain applications of topic matching where a pre-defined topic list needs to be referred to.

Given this discussion on various natural language processing techniques, it is important to establish, wherever possible, a metric for evaluating the performance of natural language processing systems.

Performance Evaluation

One of the most notable, and widely used, measures for evaluating classification tasks (like named entity recognition and relation extraction) are by computing a model's precision, recall, and F1 score. The following list defines each of these metrics:

- **Precision:** Refers to the fraction of predictions that are relevant to the truth sample. I.e. it is the ratio of total predictions generated by a model that can be considered as 'correct'.
- **Recall:** Refers to the fraction of 'correct' instances from a 'truth' set that were reproduced by the model. I.e. in supervised training of a classification model, a 'truth' dataset is what the model will train on and the higher occurrence of distinct 'truth' examples in the model's output results in a higher recall score.
- **F1:** The harmonic mean of precision and recall.

As is typical in reporting, these scores fall between the 0-1 range with values closer to one indicative of higher performance in that metric. Considering true positives/negatives, false positives, and false negatives, evaluating these metrics is straightforward:

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3.1)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} \quad (3.2)$$

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3.3)$$

However, these metrics are best applied when a model's output follows a strict binary positive/negative format. In natural language processing classification tasks however, the outputs can be more nuanced, and an alternative performance metric would provide a better indication of a model's performance.

Adapted PRF Metrics (MUC-5)

The metrics established in [51] are one such adaptation better suited for more ambiguous classification tasks. Here, the true/false positive/negative variables are replaced with the following scoring categories:

<input type="checkbox"/> Correct	response = key
<input type="checkbox"/> Partial	response \equiv key
<input type="checkbox"/> Incorrect	response \neq key
<input type="checkbox"/> Spurious	key is blank and response is not
<input type="checkbox"/> Missing	response is blank and key is not
<input type="checkbox"/> Noncommittal	key and response are both blank

Figure 7: Scoring categories established in [51].

The most notable metrics in this addition are that of the ‘Partial’ and ‘Spurious’ categories with the former referring to model predictions that are partially correct and the latter referring to model predictions that are neither correct nor incorrect (i.e. they have no reference figure in a ‘truth’ set). As will be discussed in the following chapter, these added categories provide a more thorough representation of a model’s performance when said model’s outputs are not simply binary in nature.

Research Question

Because of the above discussions regarding science traceability and relevance, natural language processing tools and techniques, and the potential benefit natural language processing can bring to mission concept evaluation efforts, the work reported in this thesis will aim to:

- **Determine the utility** of natural language processing in the domain of mission concept evaluation processes by:

- **Determining** the capabilities of natural language processing in extracting a mission concept's scientific goals, objectives, and requirements and;
- **Evaluating** those data products against the needs of the scientific community so as to assist with generating a recommended portfolio of mission concepts.

The Astrophysics Decadal process is a ripe area to employ natural language processing as alluded to in Chapter 2. Furthermore, and through discussions with NASA-affiliated researchers, ambitions to employ AI/ML towards the decadal process are pronounced and have been drivers of this NSTGRO-supported work. As such, we believe we can provide another substantial use case of natural language processing in yet another systems engineering domain. The following chapter will discuss our contribution to this gap by discussing the methods employed in AstroNLP and the results of our contribution.

CHAPTER IV

NATURAL LANGUAGE PROCESSING FOR MISSION CONCEPT EVALUATION

Science Traceability Extraction

All the work conducted in this thesis was captured in a python-based tool that deployed the above-mentioned natural language processing techniques towards science traceability extraction and relevance assessment on astrophysics-based space mission concepts. This tool, called AstroNLP, coupled methods of PDF document parsing, tokenization, NER and relation extraction through the use of specialized transformer models, and analyzed term frequencies against a constructed knowledge base of concepts to evaluate scientific relevance both at a mission and portfolio level. The figure below represents the architecture of the tool:

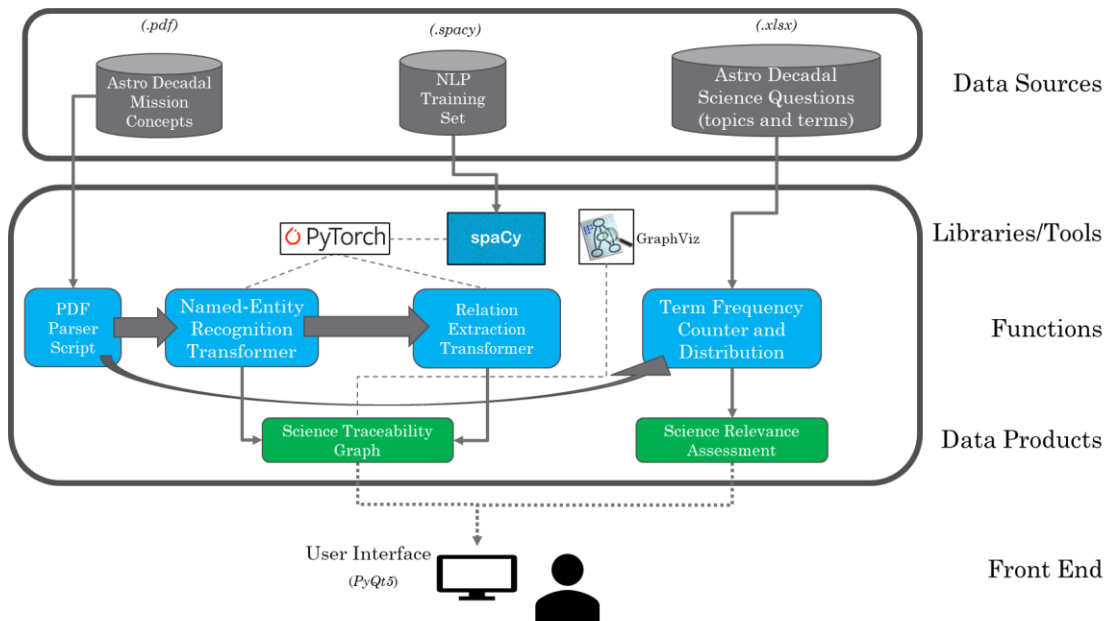


Figure 8: AstroNLP software architecture showing functions, data sources, process/data flows, and accompanying open-source libraries/tools.

For matters of clarity, the above figure can be reduced to its mere functional work-flow showing the inputs to the tool (mission concept documentation) as well as its outputs (e.g. science traceability graphs, and scientific relevance charts listed in green). The following figure details this functional flow:

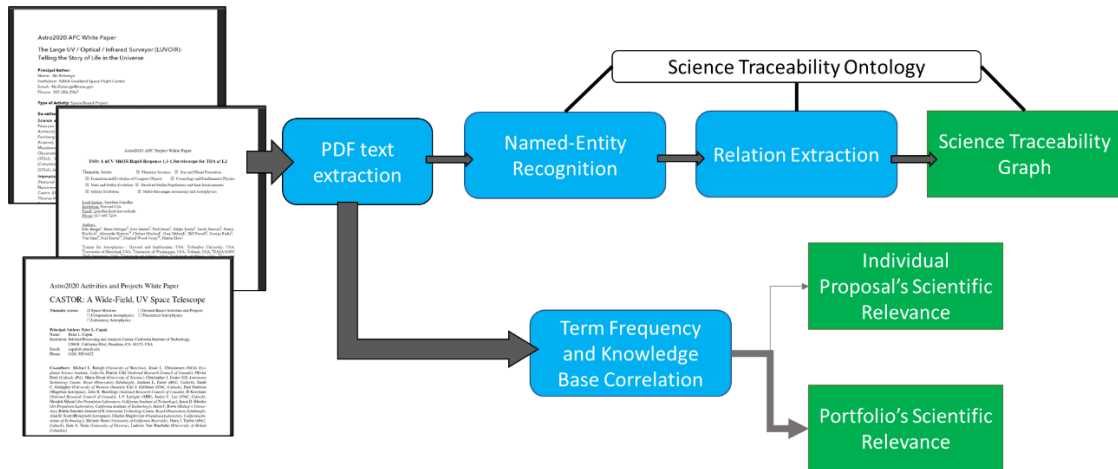


Figure 9: Functional work-flow showing the processing pipeline of the AstroNLP tool.

With the above architecture visualized, the following sections will focus on describing the individual elements contained within the functional flow and associated processes/libraries/tools utilized to support those functions.

Science Traceability Graph Ontology

As discussed in Chapter 3, it is useful to establish some form of semantic structure to our processing pipeline to guide the named-entity recognition and relation extraction transformers. This is particularly necessary when building custom annotations in a training dataset as you are required to establish what are your entity and relation types. For AstroNLP, we utilized the existing taxonomical structures established in [4, 7] as well as the Semantic Sensor Network/Sensor, Observation, Sample, Actuator Ontologies [52], all of which have been discussed in a prior publication of this work [10]. A summary of the entity types is provided below along with associated definitions and examples:

Table 3: List of all 10 entity types with associated descriptions and examples as seen in [10].

Entity Type	Entity Description	Examples
MISSION	Any word or phrase containing the name and/or acronym of the mission	Hubble, JWST, LUVOIR, The Spitzer Telescope
MISSIONPARAMETER	Any word or phrase that can be considered as an attribute describing the mission	Orbit, inclination, lifetime, cost
MISSIONPARAMETERVALUE	The quantitative or qualitative value of a MISSIONPARAMETER entity	1400 km, \$4M, 3 years
SCIENCETHEME	Any word or phrase defining, implying, or relating to a scientific topic and/or feature of interest	Black holes, early universe, galaxy, Hawking radiation
SCIENCEACTION	Any word or phrase describing an activity or set of activities a mission or instrument will perform to generate a data product and/or achieve a science goal	High-contrast direct observations, complete full-sky survey
INSTRUMENT	Any word or phrase defining an instrument contained within the mission	Spectrograph, telescope, NIR coronagraph
INSTRUMENTPARAMETER	Any word or phrase that is considered as an attribute describing an instrument	Field of view, aperture, diameter, angular resolution

INSTRUMENTPARAMETERVALUE	The quantitative or qualitative value of an INSTRUMENTPARAMETER entity	180 deg, 6 meters
OBSERVABLEPARAMETER	Any word or phrase defining an attribute of a SCIENCETHEME that the mission/instrument aims to measure	Stellar brightness, radial velocity, spectral range
OBSERVABLEPARAMETERVALUE	The quantitative or qualitative value describing the required value for an OBSERVABLEPARAMETER entity	100,000 observations, 70% of the sky each orbit, 5 – 7 keV

As listed in the above table, there are several notable entity types that are influenced by prior works and will be discussed now. The ‘SCIENCETHEME’ and ‘SCIENCEACTION’ entity types were heavily influenced by similar instances in the P-STAF architecture [8] and also discussed in [10]. Additionally, the ‘INSTRUMENT’, ‘INSTRUMENTPARAMETER’, and ‘INSTRUMENTPARAMETERVALUE’ and the parallels along the ‘MISSION’ entity types (and related) drew motivation from the original STM taxonomy (also reported in [10]). Furthermore, the ‘OBSERVABLEPARAMETER’ and ‘OBSERVABLEPARAMETERVALUE’ entities draw motivation from SSN/SOSA’s “Observable” and “ObservableProperty” entity types as discussed in [10]. Finally, as mentioned in [10], it is important to recognize that building a universally accepted science traceability taxonomy is a challenge in-of-itself

that must balance the views and perspectives of all stakeholders involved. Thus, this entity taxonomy merely represents one such possible guideline for formalizing the science traceability nomenclature.

To round out the construction of this science traceability ontology, the previously established entities must also have some form of relations associated with them so as to create a graph ontology with nodes (entities) and edges (relations). The following figure represents this final science traceability ontology utilized in this work:

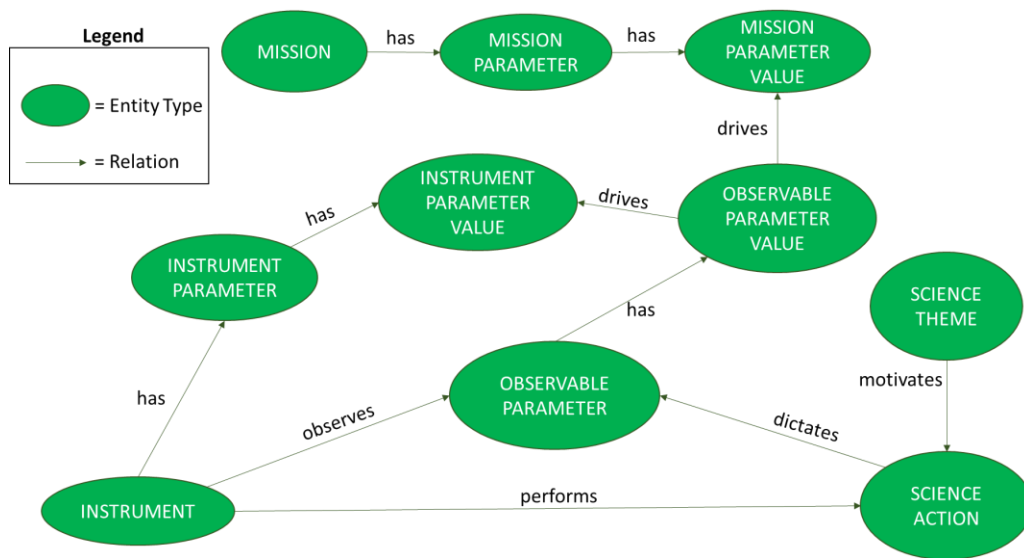


Figure 10: Science traceability graph ontology governing the AstroNLP system as seen in [10]. This governing ontology serves as the guiding template for annotations.

With the science traceability ontology defined, the following sections will discuss the specific processing pipeline employed in AstroNLP starting with the portion of the pipeline associated with extracting science traceability graphs.

Document Text Extraction

The first element of the pipeline employs a custom-built python script aimed at extracting the raw text from PDF based documentation (mission concept proposals). This script employs the py-pdf-parser open-source library [53] which builds off the fundamental framework of the pdfminer-six library [54] to extract text from PDF documents. This works through a form of object character recognition (OCR) where each individual character in the document is parsed and then grouped together in so called ‘elements’ through a series of heuristics [54]. The script then takes these elements and targets the related ‘science’ and ‘engineering’ sections through an added layer of filtering to target the most relevant portions of the document. In the case of decadal proposals, these sections typically include the ‘Key Science Goals and Questions’ and ‘Technical Overview’ sections of the document. However, not all documents follow the same header labels but rather have headers with slight deviations to the base nomenclature. As such, the parser utilizes a likelihood function that aims to index where the most likely relevant headers are located to help guide the extraction. Finally, the parser script filters any extracted ‘element’ that falls short of a minimum character threshold so as to filter out any unnecessary noise within the document (e.g. page numbers, headers, and footers).

Following text extraction, the raw text is then fed through the transformer portions of the pipeline. Prior to transformer processing, the text passes through one final filter aimed at removing any non-ASCII characters and replacing common Greek

characters with their spelled-out form. The next two sections will now discuss the transformers that perform the NER and relation extraction tasks.

Named-Entity Recognition Transformer

The NER transformer is built and trained through the expansive python library spaCy [33]. There exist several other open-source libraries capable of performing entity extraction including the Java-based CoreNLP [55] and OpenNLP [56] as well as the python-based NLTK [32]. However, due to the thorough online documentation, application programming interface, and recent support for state-of-the-art transformer models in spaCy 3.0, the python-based spaCy library was chosen for this task.

A pre-trained transformer model based in PyTorch, and pulled from Huggingface's vast repository of transformer models, is used for the named entity recognition task. Specifically, we selected the 'allenai/scibert_scivocab_cased' model [39] as our pretrained transformer and with spaCy, fine-tuned the transformer through our custom developed training data set. SciBERT was ultimately chosen due to its relevant pre-training on a scientific corpus and as such, is relevant for application on astronomy/astrophysics-based proposals. Further model parameters for training are summarized below:

Table 4: Named entity recognition model parameters.

Model Parameter	Value
model_type	bert
vocab_size	31116
batch_size	128
training.optimizer	Adam.v1
max_epochs	30
tokenizer	spacy.tokenizer.v1

As is the case for both the NER and relation extraction transformers, we kept the epoch count high in order to produce two transformer models at the end of training (a ‘low-epoch’ and ‘high-epoch’ model where the low-epoch model is the ‘best scoring’ model determined via spaCy’s training API). This was done as preliminary analyses regarding the end-to-end performance of the processing pipeline significantly varied based upon type of model used (i.e. a low-epoch model may not produce enough entities to reasonably populate a STG whereas a high-epoch would produce a substantial pool of entities for an STG, some of which may simply be noise). Example outputs from both case models will be discussed later in this chapter.

Finally, upon successful extraction of entities through the NER transformer, the entity pools go through final slew of POS-based heuristics as a post cleaning step. Specifically, all ‘Science Action’ entities must contain a ‘Verb’ or ‘Noun/Proper Noun’ token while also restricting any entities containing ambiguous end tokens (e.g.

‘Adjectives’, ‘Coordinating Conjunctions’, and ‘Auxiliaries’ to name a few). Similarly, all ‘Science Theme’s must contain a ‘Noun/Proper Noun’ POS tag and holds the same end token rule mentioned previously. We also restrict all entities to not exceed a maximum token length of 10 with specific restrictions on ‘Parameter Value’ entity types that need to contain at least 2 tokens (and also must have a ‘Numeral’ token present).

Relation Extraction Transformer

The relation extraction transformer follows a very similar parameter architecture to that of the named-entity recognition transformer. This model also uses SciBERT as the base pre-trained model and is also employed, and trained, through spaCy’s API. The model parameters for the relation extraction model are provided below:

Table 5: Relation extraction model parameters.

Model Parameter	Value
model_type	bert
vocab_size	31116
batch_size	128
training_optimizer	Adam.v1
max_epochs	100
tokenizer	spacy.tokenizer.v1

In addition to the relation extraction transformer model itself, a custom script is added to the pipeline to add filtering rules to all predicted relations. Specifically, each relation generated by the model also comes with a confidence score (valued between 0 and 1). Any predicted relation that comes with a confidence score below an adjustable threshold value will be filtered out of the pipeline. Further, a series of ‘if’ statements is applied to hardcode the restrictions enforced by the governing ontology mentioned previously. It should be noted that although the training data sets follow the entity-relation structure provided by the ontology, the model is still stochastic in nature and ‘false’ relation predictions can still emerge from the model. For those reasons, the ‘if’ statements are included as another layer of filtering.

The next section will discuss the training procedure and current scope of the training data set for both transformer models employed.

Tool Training

All annotations were developed through an online-based annotation tool, UBIAI [57]. This tool was selected for its intuitive user interface, capabilities for providing both entity and relation training data, and low-cost. The following figure illustrates the graphical interface provided by this online tool:

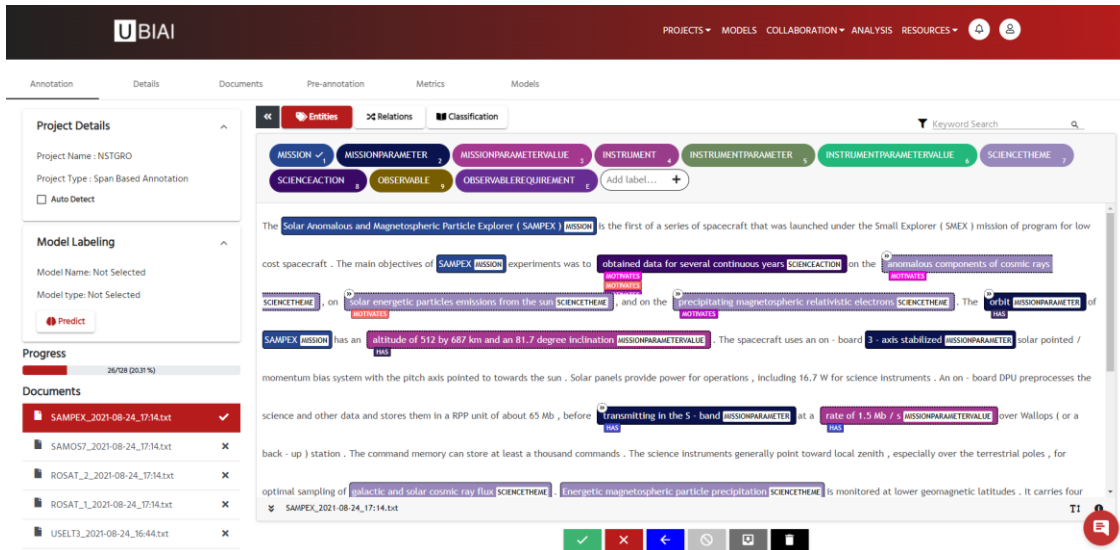


Figure 11: UBIAI's graphical user interface when viewed from the document annotation portal [57].

This online tool holds several features most notably the ability to add custom entity and relation types, upload documentation for annotation, download annotations in various formats, add contributors to support data development, and the ability to auto-annotate documents through its ‘learned’ annotator (i.e. its own annotation model learns off of the user’s provided annotation data and can be used to create predictions of its own on user-uploaded documentation). For our case, we chose to manually annotate all documentation as the auto-annotator capability has only recently become available.

The source material used to build our training data set comes from two corpora: the NASA Space Science Data Coordinated Archive’s (NSSDCA) spacecraft database [58], and the Astrophysics Data System (ADS) [59]. Both contain several descriptions and links to many prior and proposed astrophysics-based space missions with the latter having direct access to many proposal documentation. In fact, ADS is the prime

repository for all concept proposals submitted to the Astrophysics Decadal Survey. With these corpora, we had access to a rich data pool for generating annotations for training.

With the UBIAI tool and these data sources, we created the following training dataset for both entity and relation annotations:

Table 6: Training data size across all entity and relation types acquired over the course of 12 months.

Named Entity Recognition Training Data Size	
Label Type	Number of Examples
MISSION	484
MISSIONPARAMETER	143
MISSIONPARAMETERVALUE	141
INSTRUMENT	291
INSTRUMENTPARAMETER	255
INSTRUMENTPARAMETERVALUE	187
SCIENCETHEME	383
SCIENCEACTION	273
OBSERVABLEPARAMETER	202
OBSERVABLEPARAMETERVALUE	133
Relation Extraction Training Data Size	
Label Type	Number of Examples
HAS	626

PERFORMS	173
OBSERVES	121
MOTIVATES	220
DICTATES	90
DRIVES	96

Due to the variances in example amount across all entity and relation types, biases do exist in our models. Whilst we have attempted to balance these variances during our annotation sprints, it is important to note the difficulty of developing an evenly distributed training dataset, especially when considering the nature of our data sources. To provide a highlight of the nature of these variances, the following tables summarizes the average and standard deviation of the character lengths for each entity type:

Table 7: Profile analytics across all ten entity types.

Entity Type	Average # of Characters	Standard Deviation
MISSION	10.886	12.920
MISSIONPARAMETER	13.154	9.239
MISSIONPARAMETERVALUE	13.823	10.993
INSTRUMENT	23.330	12.249

INSTRUMENTPARAMETER	15.580	6.752
INSTRUMENTPARAMETERVALUE	13.369	8.603
SCIENCETHEME	21.201	10.582
SCIENCEACTION	39.498	18.710
OBSERVABLEPARAMETER	18.540	11.095
OBSERVABLEPARAMETERVALUE	19.564	10.597

The following section will discuss the graph visualization performed post-transformer processing.

Graph Generation and Visualization

The final step in the processing pipeline aims to formulate all the entities and relations into a visual artifact for review by a user. Here, we employ the GraphViz [60] engine to essentially print the entities and relations as nodes and edges respectively. To do this, we constructed a script that uses a python-based GraphViz API that can communicate with the GraphViz software to structure and format the graph in a way that not only captures the extracted entities and relations, but also orders the graph by various sub-groups detailing science themes, science actions, observable requirements, instrument requirements, and mission requirements.

Upon completion of the science traceability extraction task, we expect the model to produce the following output (NOTE: this is a manually constructed output as reported in [10] on a textual sample obtained from the Cosmic Dawn Intensity Mapper concept study [5]):

1 EXECUTIVE SUMMARY

The Cosmic Dawn Intensity Mapper (CDIM) will transform our understanding of the era of reionization when the Universe formed the first stars and galaxies and UV photons ionized the neutral medium. CDIM goes beyond the capabilities of upcoming facilities by carrying out wide area spectro-imaging surveys providing redshifts of galaxies and quasars during reionization as well as spectral lines that carry crucial information on their physical properties. CDIM will make use of unprecedented sensitivity to surface brightness to measure the intensity fluctuations of reionization on large-scales to provide a valuable and complementary dataset to 21-cm experiments. The concept is an 83-cm infrared telescope equipped with a focal plane of 4×2048^2 detectors capable of $R = 300$ spectro-imaging observations over the wavelength range of 0.75 to 7.5 μm using Linear Variable Filters (LVFs). The large field of view of 7.8 deg² allowing efficient wide area surveys, and instead of moving instrumental components, spectroscopic mapping is realized through a shift-and-stare strategy through spacecraft operations. CDIM design and capabilities focus on the needs of detecting faint galaxies and quasars during reionization and intensity fluctuation measurements of key spectral lines, including Lyman- α and H α radiation from the first stars and galaxies. The design is low risk, carries significant science and engineering margins, and makes use of technologies with high technical readiness level for space observations.

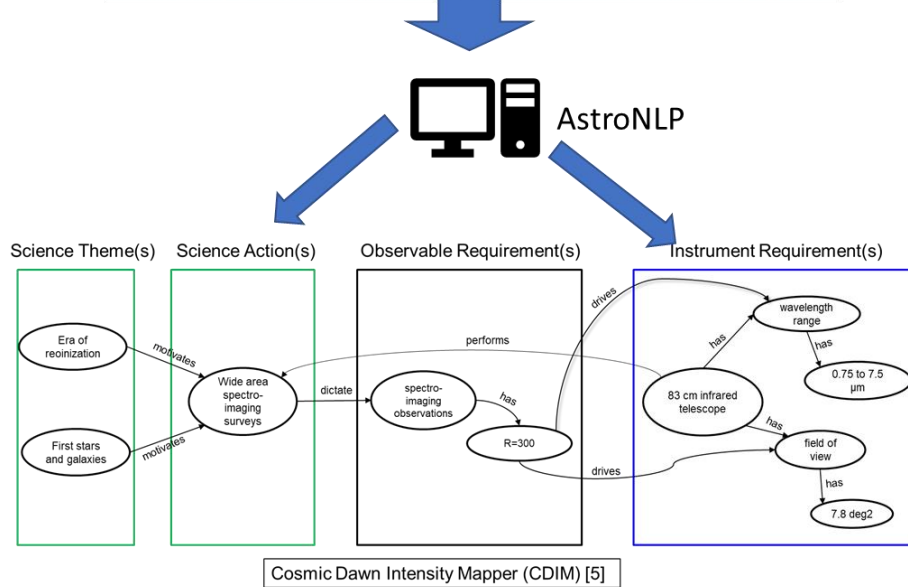


Figure 12: Illustration of the input and output expectations for the AstroNLP tool.

It is important to note that the actual tool will process documentation that are multiple pages in length (typically between 10 and 20 pages). As such, the above example is merely an apportioned representation to illustrate the end product.

Science Relevance Assessment

Science Panels and Questions Knowledge Base

To address the science relevance question, we utilize a form of term frequency and topic matching by tokenizing the input text acquired from documentation and

comparing extracted noun-chunks with topics inside a knowledge base. This knowledge base, constructed in excel, follows the structure seen in the decadal's science panels and science questions as reported in Chapter 2. In this excel document, the rows refer to each individual science question (24 total) given in the decadal survey. Each question is also given a unique identifier, and a corresponding sheet prescribes the set of terms related to that science question. This knowledge base was constructed manually based upon the perceived importance of certain themes (green), spectra (red), instrument/technique (black), and parameters (blue) provided in the appendices of the decadal report. For reference, the two sheets of the excel document are provided in Appendix A.

The following section will discuss in detail, the noun chunking and topic matching script used to correlate extracted tokens with those given in the knowledge base.

Noun-Chunking and Term Frequencies

In relation to science traceability extraction, the effort to process the text in a manner suitable for topic matching is much more simplified. No training was necessary for this portion of the work as publicly available noun-chunking models were used (specifically spaCy's en_core_web_sm model).

To generate the relevance charts, we first take the inputted textual chunks from the mission concept documentation and extract all noun chunks via the en_core_web_sm model. We then correlate these noun chunks with the topics listed in the knowledge base in order to generate a 'number of hits' per term for each and every science question (a

‘term hit’ simply refers to the occurrence of a term in a noun chunk). The results for all science questions, and individual distributions per science question across its respective topic terms, are then plotted in a histogram format to generate topic/term distributions (similar to that shown in Chapter 2). For multi-mission analysis, the textual chunks are concatenated with the related missions contained in the mission ‘portfolio’ and the outputs are generated via the same process.

The following section will discuss the layout of the tool’s graphical user interface, the performance metrics for both transformer models, as well as provide output examples for several mission concept proposals in addition to concatenated results for combinations of said missions.

Application to the Astrophysics Decadal Survey

Tool Graphical User Interface

The following figure portrays the entire visual field of AstroNLP’s graphical user interface (the GUI is built through the PyQt5 [61] python library):

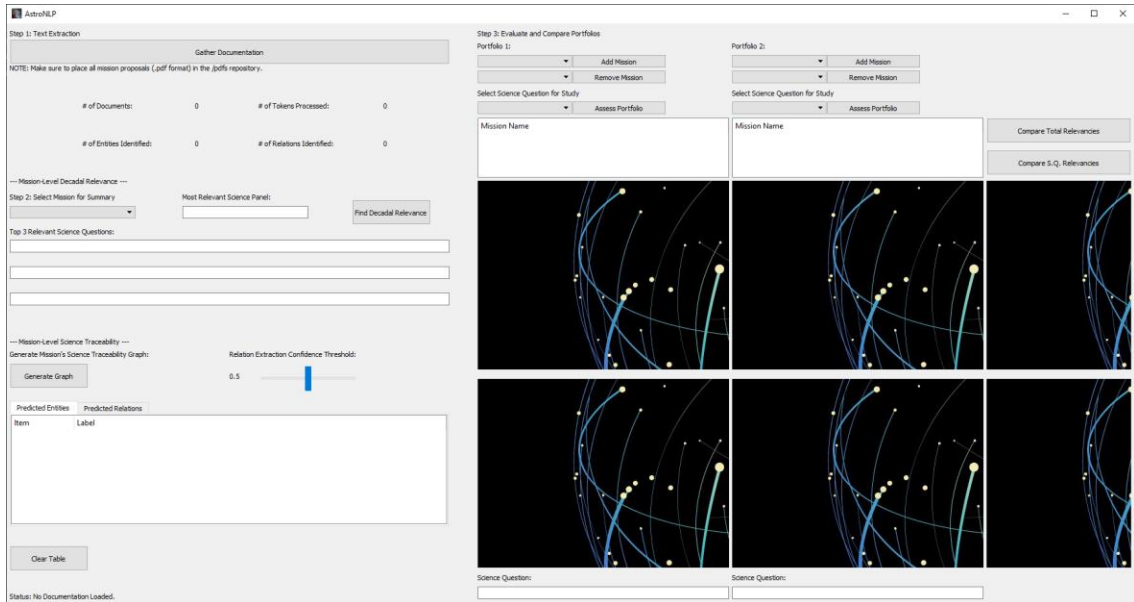


Figure 13: AstroNLP's graphical user interface.

Visually, the tool's interface can be subdivided into four primary subregions: 1) the repository metrics region, 2) the mission-level relevance assessment region, 3) the graph generation region, and 4) the mission and portfolio relevance assessment region.

In region 1, the processing of documentation starts through the activation of a single widget (i.e. pressing 'Gather Documentation'). Any PDF documents contained within the local repository are then queued and text is extracted. The following figure shows this region after loading three proposal documents:

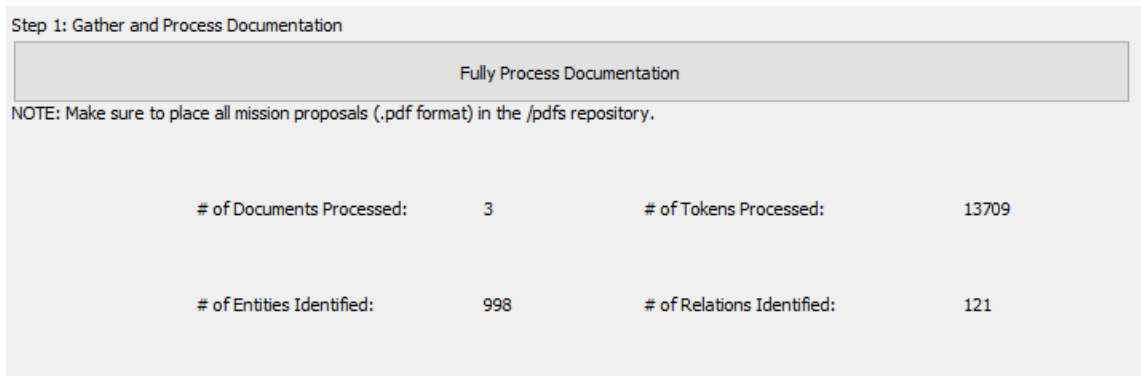


Figure 14: Region 1’s document metrics across the repository of mission concepts. Here, the total number of documents, entities, tokens, and relations are provided for visual inspection to the user.

In region 2, a specific mission concept can be evaluated for relevancy against the knowledge base’s scientific questions. Here, a specific mission from the repository can be selected and then the output returns the most relevant science panel and top three related science questions for that mission (via the same process discussed in the prior section). An example output for the LUVOIR mission concept is provided below:

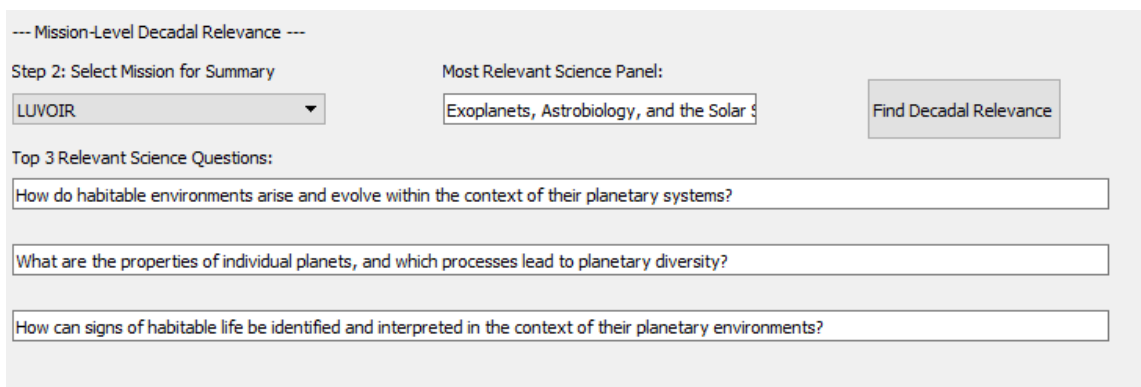


Figure 15: Region 2 of the AstroNLP tool showing specific science panel/question relevancies for the LUVOIR concept.

In region 3, graph generation takes place. After processing, all entities and relations are stored in the local cache of the tool, and it is here where a specific mission's science traceability graph can be produced. The open list shows all entities and valid relations extracted during the initial processing period and can be viewed individually. An example output (excluding the science traceability graph itself) for the LUVOIR mission concept is shown below:

--- Mission-Level Science Traceability ---
 Generate Mission's Science Traceability Graph:

Generate Graph

Predicted Entities Predicted Relations

Entity	Instance	Relation	Entity	Instance	Confidence
SCIENCETH...	galaxies	MOTIVATES	SCIENCEACTION	detailed observ...	0.9128059
SCIENCETH...	exoplanets	MOTIVATES	SCIENCEACTION	detailed observ...	0.9002169
SCIENCETH...	solar system bo...	MOTIVATES	SCIENCEACTION	provide near-fly...	0.66046685
SCIENCETH...	gas giant exopl...	MOTIVATES	SCIENCEACTION	measure the at...	0.99475914
SCIENCEAC...	read the fingerp...	DICTATE	OBSERVABLE	wavelengths	0.5124644
SCIENCETH...	matter	MOTIVATES	SCIENCEACTION	access to a ra...	0.62712234
SCIENCETH...	matter	MOTIVATES	SCIENCEACTION	read the fingerp...	0.9973327
MISSIONPA...	prime mission	HAS	MISSIONPARA...	5-year	0.59513366
MISSIONPA...	lifetime	HAS	MISSIONPARA...	5-year	0.7405785

Clear Table

Figure 16: Region 3's science traceability graph's metrics for a specific mission concept (LUVOIR in this example). Notice that both the entity and relation lists can be viewed by selecting the appropriate tab.

In region 4, a mission or portfolio can be assessed against specific science questions provided in the decadal in order to determine the term/topic distributions for a selected science question. This is done via the same process discussed in the recent

section and can be evaluated for any combination of mission concepts provided in the repository. Additionally, two portfolios can be compared against one another in regards to their perceived scientific impact across all science questions and/or select science questions. An example output is provided below:



Figure 17: Region 4's relevance assessment panel. Here, the histogram plots can be viewed across two portfolios under comparison detailing the impacts these portfolios have on all/select science questions (science questions are listed via their ID number and printed after assessing the 'portfolio').

Further examples of specific science traceability graphs and relevance assessments for various mission concepts will be discussed later in this chapter.

System Performance

For both transformer models, we constructed a custom python script to evaluate the baseline fine-tuned models against the testing dataset (the testing dataset is roughly 20% of the total annotation dataset and is disjoint from the training dataset). Given the performance metric equations discussed in the previous chapter (and in [51]), we follow the scoring strategy reported in a previous publication [10] and is as follows:

$$Precision = \frac{Correct\ Instances}{Correct\ Instances + Incorrect\ Instances + Spurious\ Instances} \quad (4.1)$$

$$Recall = \frac{Correct\ Instances}{Correct\ Instances + Incorrect\ Instances + Missed\ Instances} \quad (4.2)$$

The F1 score is calculated through the same equation (3.3) provided in the previous chapter. Regarding the additional scoring categories from [51] and mentioned in the previous chapter, one notable alteration we included in the script was the combination of ‘Correct’ and ‘Partial’ instances. In essence, partial instances were counted based upon their ‘coverage’ of a true/correct instance. To put it simply, when counting instances across all categories, all truly ‘Correct’ instances were given a value

of 1 and ‘Partial’ instances were given a value between 0 and 1 depending upon how much of the true instance was captured. This was done by measuring the amount of overlap that occurred for a predicted span and a ‘true’ span in a textual example.

With this evaluation, we received the following baseline performance metrics for both transformer models:

Table 8: Baseline performance metrics for both transformer models based upon gold annotations.

Transformer Model	Precision	Recall	F1
NER Transformer	0.34	0.13	0.19
Relation Extraction Transformer	0.49	0.28	0.35

Note that the above table only considers ‘baseline’ model performance which does not factor in the effects of POS filtering and ontology enforcement. By performing the same scoring procedure on the ‘enhanced’ processing pipeline (e.g. with filtering heuristics) we receive the following performance metrics:

Table 9: Full pipeline performance metrics for both transformer models. This scoring was carried out semi-automatically with relation extraction performance based off of NER model output as opposed to the gold annotations. Additionally, this scoring procedure only used ~25% of the testing data set (roughly 5% of the total training data set).

Transformer Model	Precision	Recall	F1
NER Transformer	0.96	0.71	0.81
Relation Extraction Transformer	0.58	0.27	0.37

We recognize that these metrics, at large, do not fully meet the upper percentile scoring values reported in other literature [41]. As such, various filtering rules discussed previously (e.g. POS tags and ontology enforcement) have been applied to the pipeline in an effort to improve the quality of outputted STGs. However, we also recognize that transformer models require an extensive amount of annotation examples for fine-tuning (thousands rather than a few hundred examples across entity and relation types) and thus recommend that future implementation of this work require an extensive effort in expanding the annotation data set first and foremost.

The next section will provide the results of the tool given these performance metrics by providing a look at example science traceability graphs and relevance histograms for various mission concepts.

Graph Generation and Relevance Examples

This section serves to provide what AstroNLP is capable of producing when applied to various mission concept proposals. As pretense, this section will cover three

NASA concept missions submitted to the decadal: two flagship concepts and one probe concept. All NASA concept missions submitted to the 2020 decadal survey can be found in [5]. Each subsection will be dedicated to one of these three concept missions with one final subsection discussing the multi-mission/portfolio level use case.

Additionally, as a further note, each mission concept discussed is given in the form of a PDF document. The document structure follows a loose template judicated by the decadal survey process, and each mission concept should contain the following main sections:

- Key Science Goals and Objectives:
- Technical Overview
- Technology Drivers
- Organization, Partnerships, and Current Status
- Schedule
- Cost Estimates

It should be noted that the wording for each of these section headers is not followed strictly in practice and proposal documents may also contain superfluous sections such as title pages, table of contents, author lists, and bibliographies (to name a few). More information regarding the document structure can be found [9].

Flagship: The Large UV/Optical/Infrared Surveyor (LUVOIR)

The LUVOIR concept ultimately consists of two individual observatories dubbed LUVOIR-A and LUVOIR-B. The signature science cases for LUVOIR are as follows (and available at [5]):

1 - Finding habitable planet candidates
2 - Searching for biosignatures and confirming habitability
3 - The search for life in the solar system
4 - Comparative atmospheres
5 - The formation of planetary systems
6 - Small bodies in the solar system
7 - Connecting the smallest scales across cosmic time
8 - Constraining dark matter using high precision astrometry
9 - Tracing ionizing light over cosmic time
10 - The cycles of galactic matter
11 - The multiscale assembly of galaxies
12 - Stars as the engines of galactic feedback

Figure 18: Signature science cases for the LUVOIR concept as provided in [5].

Some notable technical features of the LUVOIR mission, provided within LUVOIR’s concept proposal, are as follows:

Table 10: Notable technical design features of the LUVOIR concept as seen in [5].

	LUVOIR-A	LUVOIR-B
Parameter	Value	Value
Telescope Diameter	15 m	8 m
Prime Mission Lifetime	5 years	
Orbit	Sun-Earth L2	

Total Observation Wavelength Range	100-2500 nm
Tracking Speed	60 milliarcseconds/sec
Instruments	HDI (near-UV – near IR imager), ECLIPS (coronagraph with imaging cameras and spectrographs), LUMOS (far-UV imager and multi-resolution, multi-object spectrograph), POLLUX (point-source UV spectropolarimeter)

Upon inputting and processing of LUVOIR’s concept proposal, we retrieve the following science traceability graph:

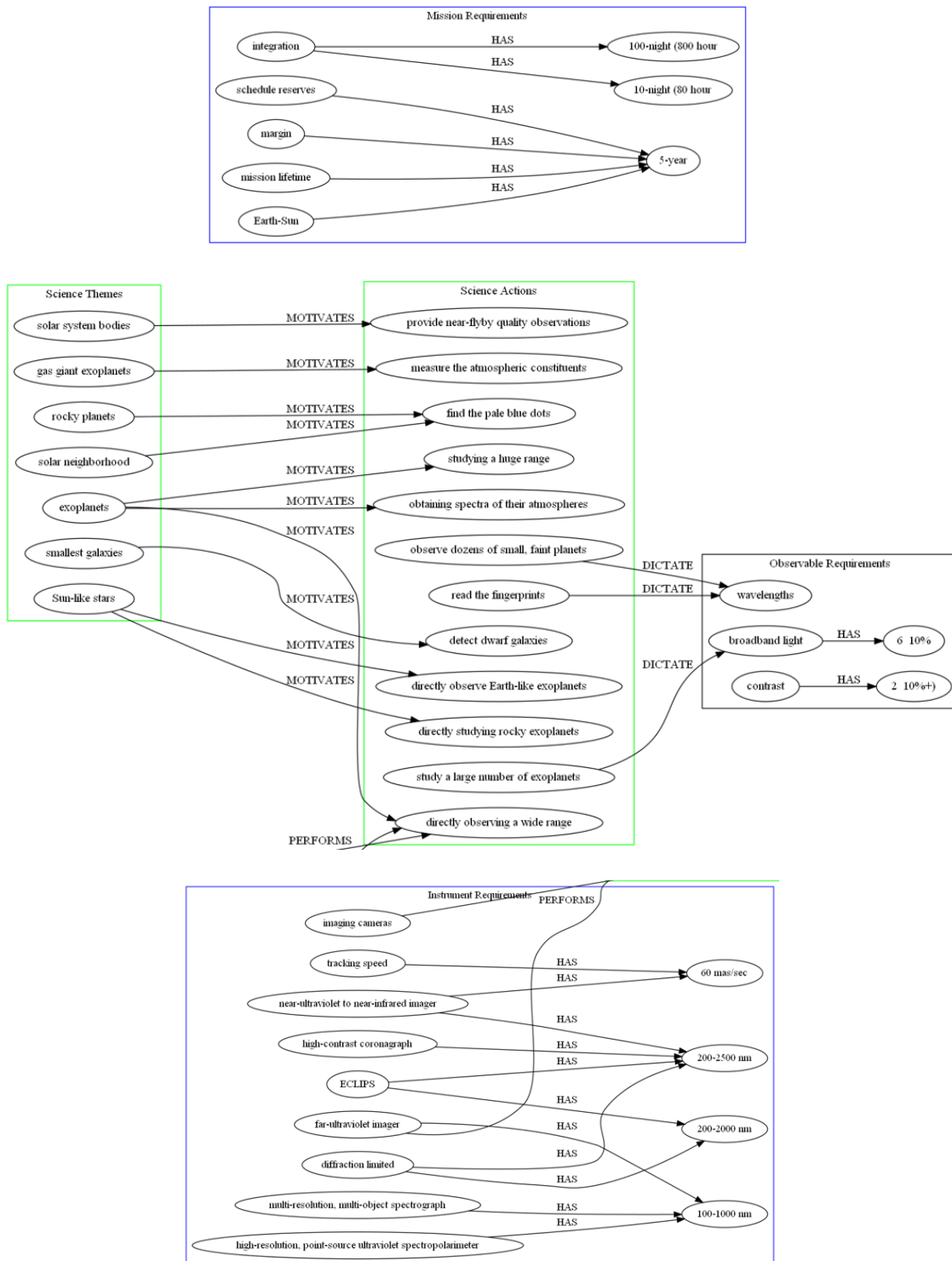


Figure 19: STG extracted from the LUVOIR concept proposal using a higher-epoch (30) NER transformer.

Looking at the scientific relevance of the LUVOIR concept across the decadal's science questions, we receive the following distribution chart:

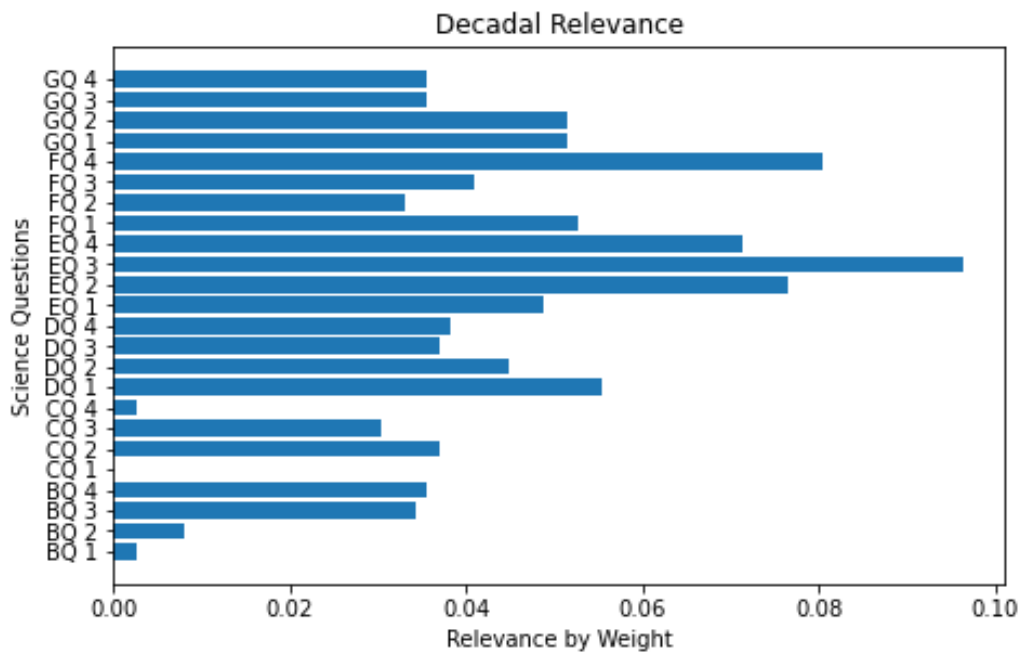


Figure 20: LUVOIR's complete relevance distribution normalized across the decadal science questions.

By this measure, and as the tool reports, the most relevant science panel for the LUVOIR concept is the 'Panel on Exoplanets, Astrobiology, and the Solar System' with the following science questions reported as being most relevant:

--- Mission-Level Decadal Relevance ---

Step 2: Select Mission for Summary

astro2020_LUVOIR.pdf

Most Relevant Science Panel:

Exoplanets, Astrobiology, and the Solar S

Find Decadal Relevance

Top 3 Relevant Science Questions:

How do habitable environments arise and evolve within the context of their planetary systems?

Is planet formation fast or slow?

What are the properties of individual planets, and which processes lead to planetary diversity?

Figure 21: LUVOIR's most relevant science panel as well as its top three most relevant science questions.

Upon specific inspection of LUVOIR's top science question, we receive the individual distribution of topics/terms for said science question represented in the following figure:

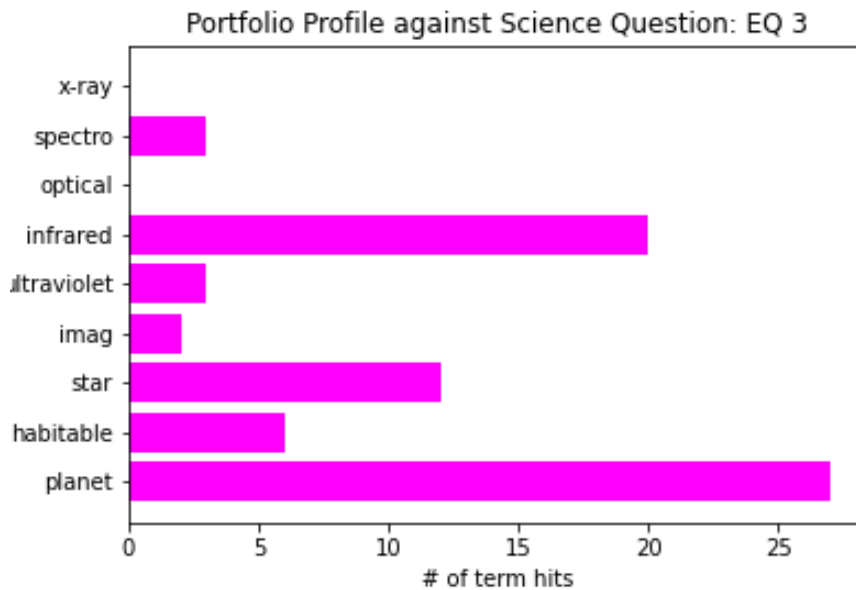


Figure 22: LUVOIR's topic/term distributions for its most relevant science question. Notice the significant relevant representation of 'infrared' and 'planet' terms.

Flagship: The Origins Space Telescope (OST) [5]

OST is one observatory looking to explore areas of galaxy formation, the origins of habitable worlds, and discover other potentially life-supporting worlds [5]. It's key science goals are as follows (and provided in [5]):




Table 1: Scientific objectives for the <i>Origins</i> Space Telescope			
NASA Goal	How does the Universe work?	How did we get here?	Are we alone?
<i>Origins</i> Science Goals	 <p>How do galaxies form stars, make metals, and grow their central supermassive black holes from reionization to today?</p>	 <p>How do the conditions for habitability develop during the process of planet formation?</p>	 <p>Do planets orbiting M-dwarf stars support life?</p>
<i>Origins</i> Scientific Capabilities	<i>Origins</i> will spectroscopically 3D map wide extragalactic fields to simultaneously measure properties of growing supermassive black holes and their galaxy hosts across cosmic time.	With sensitive, high-resolution spectroscopy, <i>Origins</i> maps the water trail from protoplanetary disks to habitable worlds.	By obtaining precise mid-infrared transmission and emission spectra, <i>Origins</i> will assess the habitability of nearby exoplanets and search for signs of life.
<i>Origins</i> Scientific Objectives	<ol style="list-style-type: none"> 1) How does the relative growth of stars and supermassive black holes in galaxies evolve with time? 2) How do galaxies make metals, dust, and organic molecules? 3) How do the relative energetics from supernovae and quasars influence the interstellar medium of galaxies? 	<ol style="list-style-type: none"> 1) What role does water play in the formation and evolution of habitable planets? 2) How and when do planets form? 3) How were water and life's ingredients delivered to Earth and to exoplanets? 	<ol style="list-style-type: none"> 1) What fraction of terrestrial planets around K- and M-dwarf stars has tenuous, clear, or cloudy atmospheres? 2) What fraction of terrestrial M-dwarf planets is temperate? 3) What types of temperate, terrestrial, M-dwarf planets support life?

Figure 23: OST's science goals and science objectives as seen in [5].

Notable technical features specific to the OST concept are provided in the table below:

Table 11: Select technical details of the OST concept [5].

Parameter	Value
Telescope Size	5.9 m
Wavelength Range	2.8 - 588 μ m
Orbit	Sun-Earth L2
Design Lifetime	5 years

Upon inputting and processing of OST’s concept proposal (of which did contain a STM), we receive the following figure:

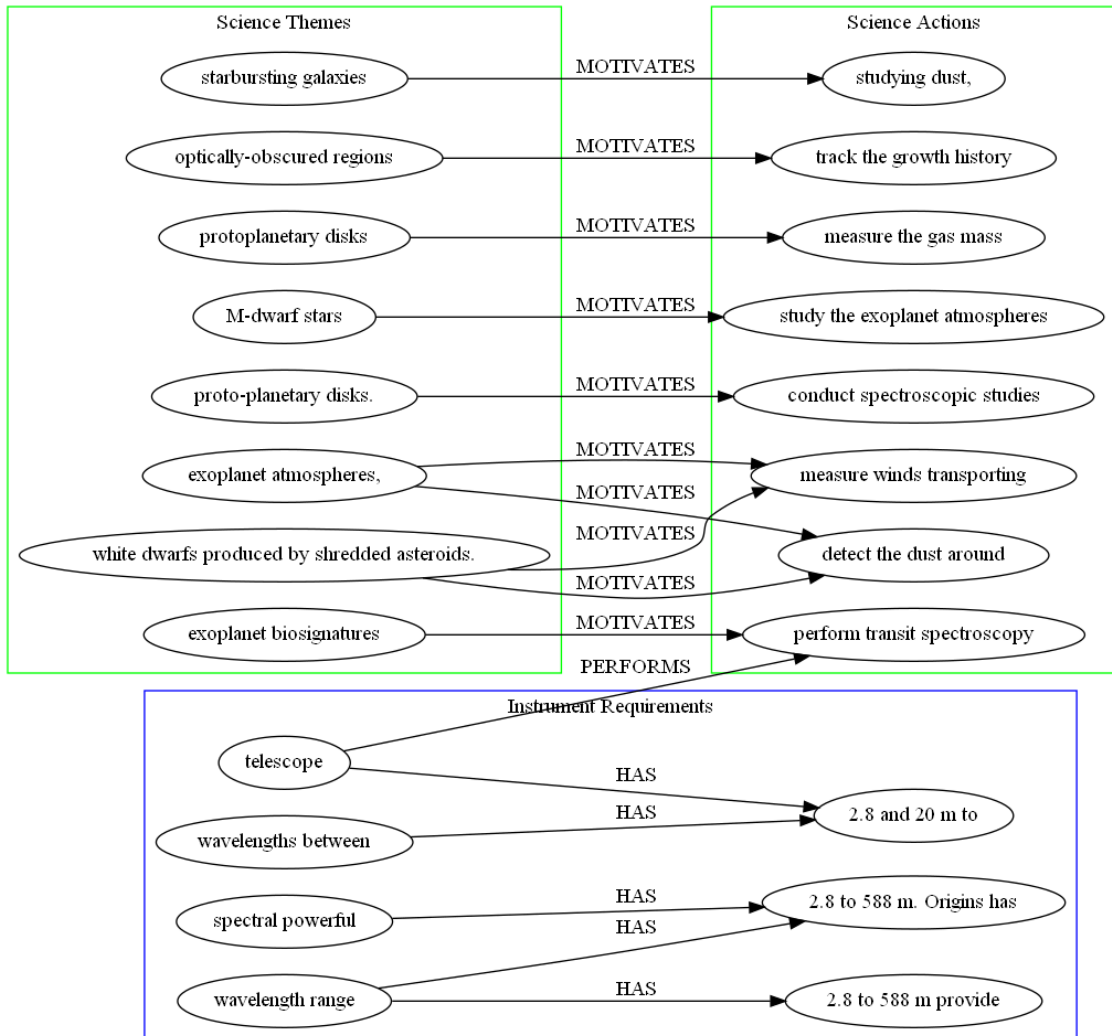


Figure 24: OST’s science traceability graph. A much larger version of this graph can be obtained if the baseline NER transformer is switched with a higher epoch model (i.e. one that went through all 30 cycles of the training data set).

As was done with the LUVOIR concept, we receive the following science question distributions for OST:

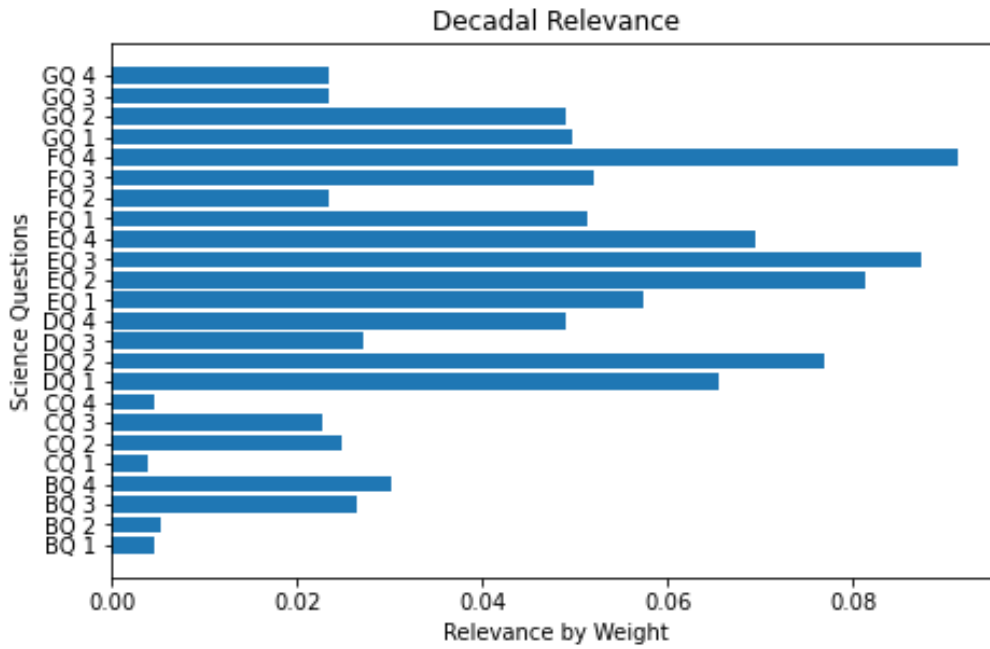


Figure 25: OST's complete relevance distribution normalized across all decadal science questions.

Again, we can also report on OST's most relevant science panel and science questions:

--- Mission-Level Decadal Relevance ---

Step 2: Select Mission for Summary

Origins

Most Relevant Science Panel:

Top 3 Relevant Science Questions:

Is planet formation fast or slow?

How do habitable environments arise and evolve within the context of their planetary systems?

What are the properties of individual planets, and which processes lead to planetary diversity?

Figure 26: OST's most relevant science panel and top three related science questions.

Upon further analysis of OST's top science question, we can attain the individual topic/term distributions for said question:

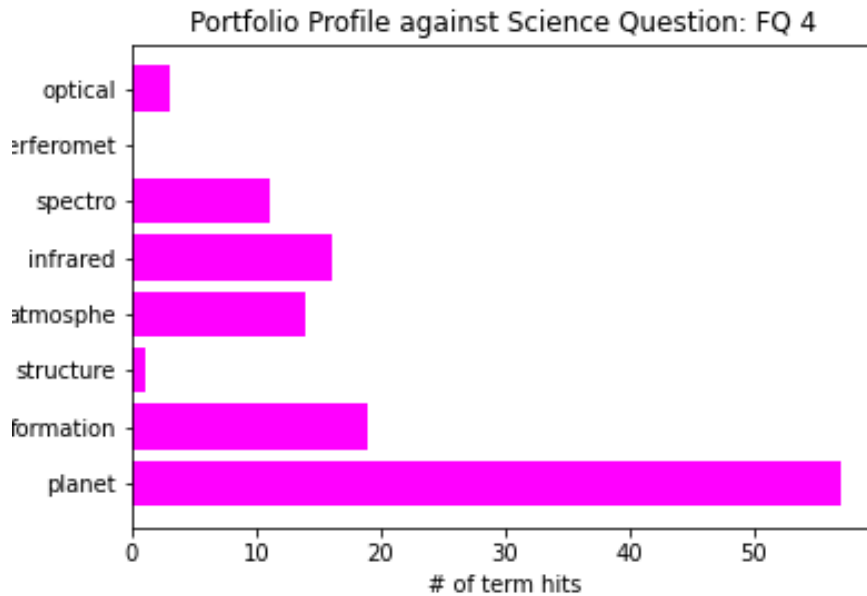


Figure 27: OST's topic/term distributions across its most relevant science question.

Probe: The Galaxy Evolution Probe (GEP) [5]

The GEP concept is a single observatory aimed at studying key concepts about star formation and supermassive black hole growth in galaxies over time [5].

Specifically, the two key science goals for GEP are listed as:

1. Map the history of galaxy growth by star formation and accretion by supermassive black holes and characterize the relation between those processes [5].
2. Measure the growth of metals over cosmic time [5].

Some notable design details of the GEP concept are reported as follows:

Table 12: Select design features of the GEP mission concept [5].

Parameter	Value
Orbit	Sun-Earth L2
Mission Duration	4 years with 46% margin
Telescope Diameter	2 m
Instrumentation	GEP-I (Imager with 23 bands covering 10-400 μm), GEP-S (Spectrometer covering the 24-42, 40-70, 66-116, and 110-193 μm range targeting select galaxies)
Total Cost Estimate	\$910M

Upon processing of GEP's mission concept document, we can attain the following STG:

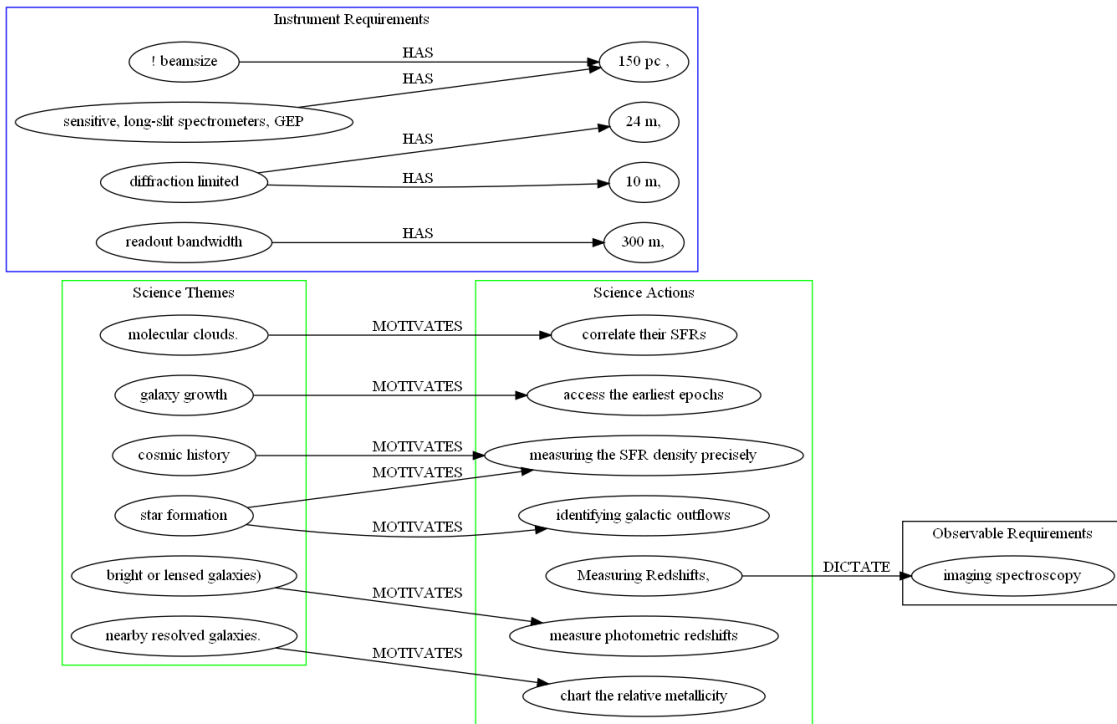


Figure 28: A portion of GEP's science theme and science action regions contained within its larger STG.

Considering GEP's relevance towards decadal science questions, we receive the following distribution:

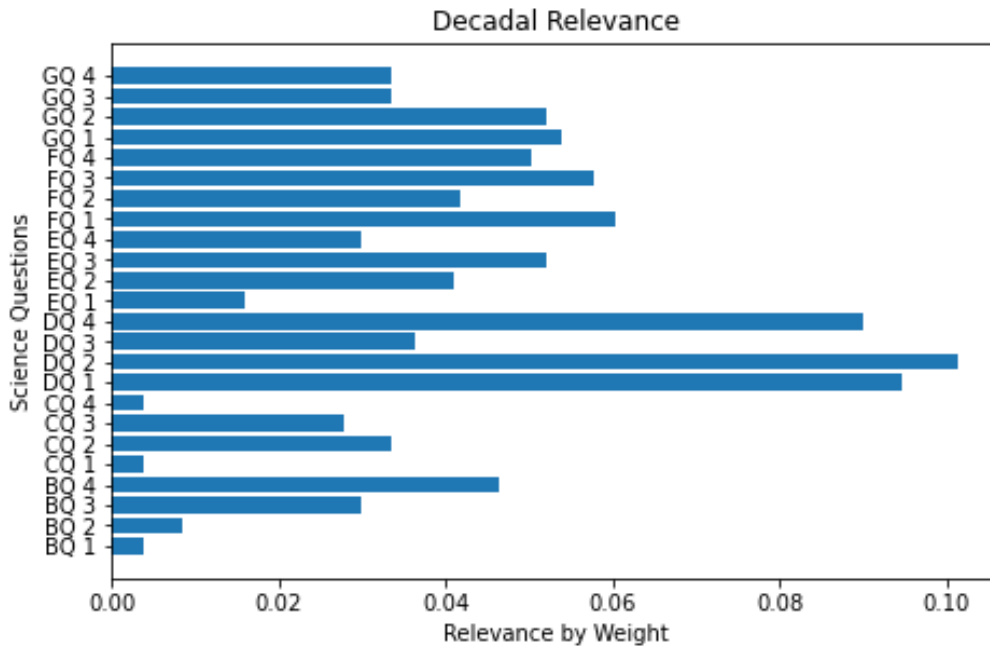


Figure 29: GEP's relevance distribution normalized across all decadal science questions.

By the distribution, we report the following most relevant science panel and top three most relevant science questions:

Step 2: Select Mission for Summary

GEP

Most Relevant Science Panel:

Galaxies

Find Decadal Relevance

Top 3 Relevant Science Questions:

How do gas, metals, and dust flow into, through, and out of galaxies?

How did the intergalactic medium and the first sources of radiation evolve from cosmic dawn through the epoch of reionization?

How do the histories of galaxies and their dark matter halos shape their observable properties?

Figure 30: GEP's most relevant science panel and top three most relevant science questions.

Upon further analysis of GEP's most relevant science question, we produce the following term/topic distribution:

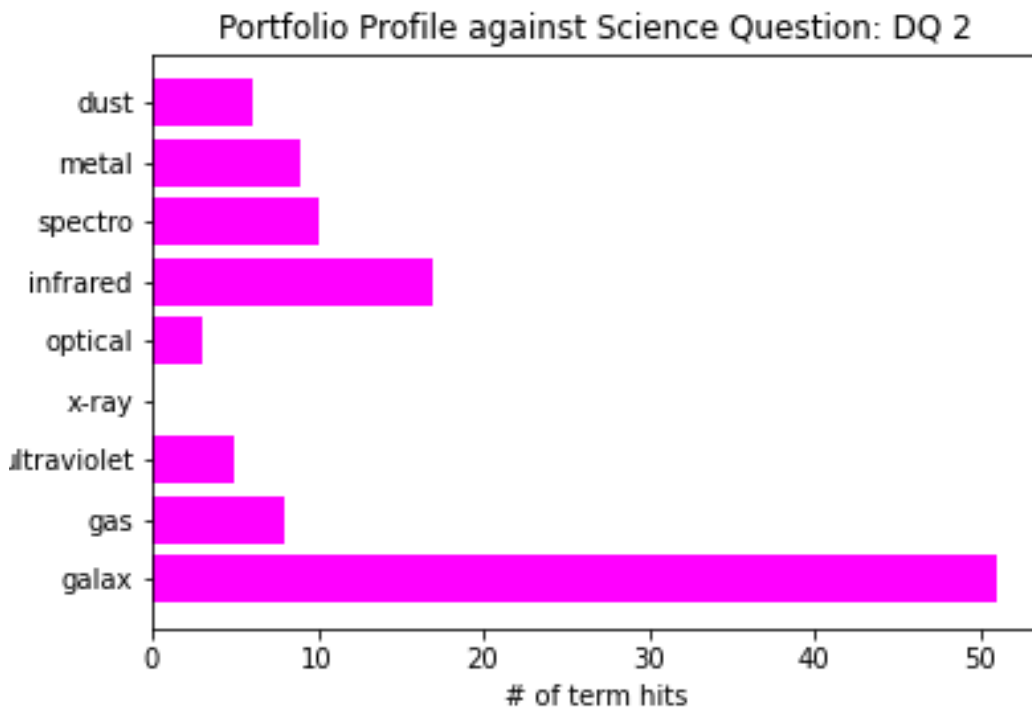


Figure 31: GEP's term/topic distribution across its most relevant science question.

In all, a total of three separate mission concept documents were analyzed (LUVOIR at 12 pages, OST at 15 pages, and GEP at 24 pages). This resulted in a processing total of ~21,000 tokens.

Portfolio-Level Analysis

Science traceability extraction is only available at the individual mission level, but an analysis of scientific relevance at the portfolio level is capable through AstroNLP. As a toy demonstration of this, consider two portfolios: one containing the LUVOIR and GEP concept (Portfolio 1) and one containing the LUVOIR and AXIS concept (Portfolio 2). AXIS stands for Advanced X-ray Imaging Satellite and is x-ray-based NASA probe concept also submitted to the 2020 Decadal [5]. In AstroNLP, and prior to analysis, these portfolios would be loaded into region 4 as illustrated in the following figure:

The screenshot displays the 'Step 3: Evaluate and Compare Portfolios' interface. It is divided into two columns for Portfolio 1 and Portfolio 2. Each portfolio has a list of missions with 'Add Mission' and 'Remove Mission' buttons. Below the mission lists, there is a 'Select Science Question for Study' dropdown menu set to 'DQ 1' and an 'Assess Portfolio' button. At the bottom of each column, a box labeled 'Mission Name' lists the loaded missions: Portfolio 1 contains 'astro2020_LUVOIR.pdf' and 'astro2020_GEP.pdf'; Portfolio 2 contains 'astro2020_LUVOIR.pdf' and 'astro2020_AXIS.pdf'.

Figure 32: Portfolio showing the 'loaded' concept portfolios ready for relevance analysis.

Assessing all these concepts provides the following total relevance profiles for each portfolio:

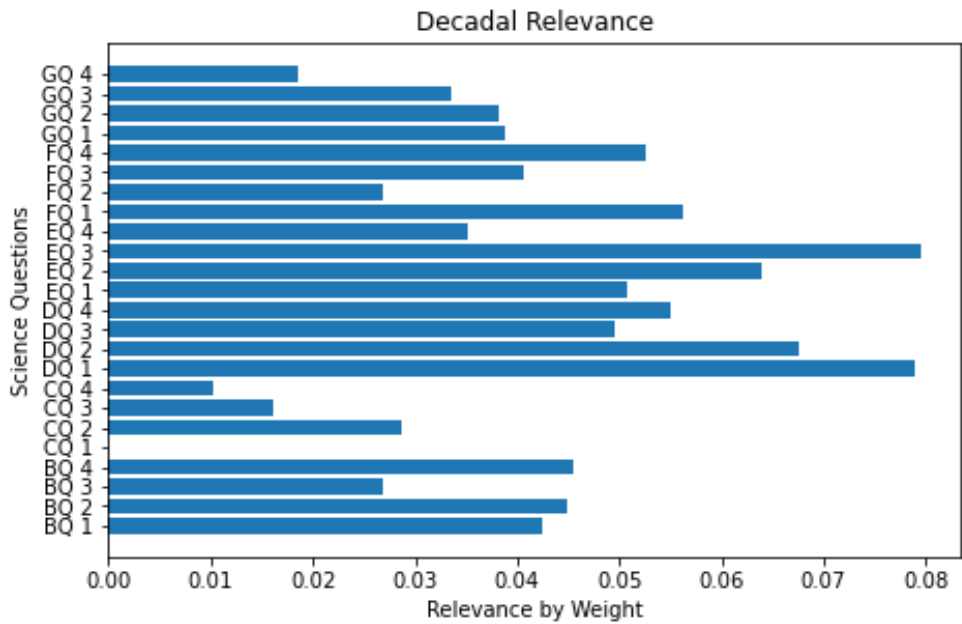
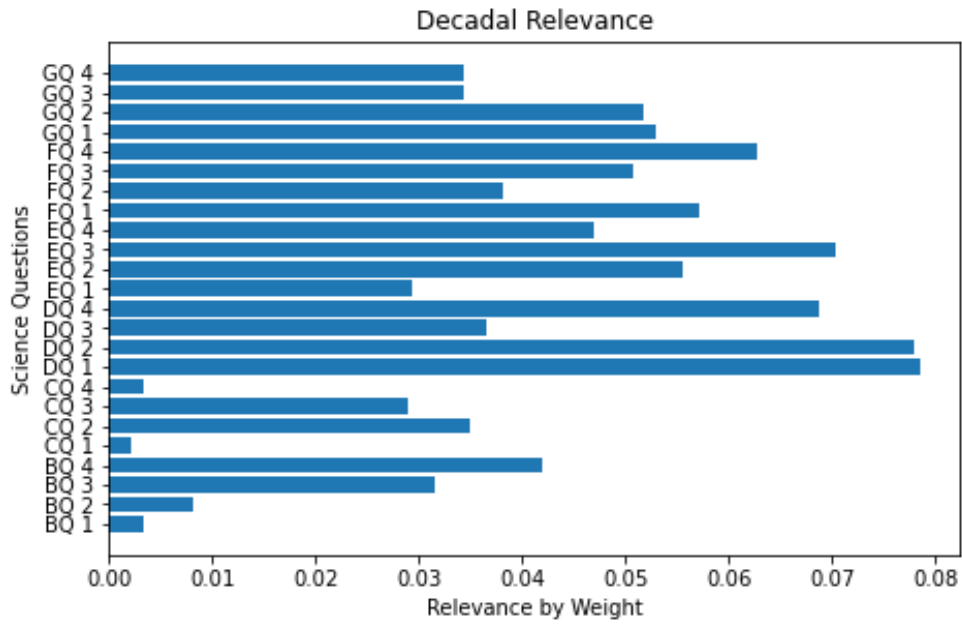


Figure 33: Each portfolio's relevance across all science questions (the upper is Portfolio 1 and the lower is Portfolio 2). This is, in essence, a concatenation of the individual relevance profiles of all mission concepts contained within either portfolio.

Upon visual inspection of these profiles, we can garner several conclusions regarding scientific impact across each question. Most notably, portfolio 2 holds a higher weight towards questions BQ-1 and BQ-2. We can also deduce that science question DQ-1 is highly relevant to both portfolios. Considering the DQ-1 question, we can attain the following term/topic distributions for both portfolios:

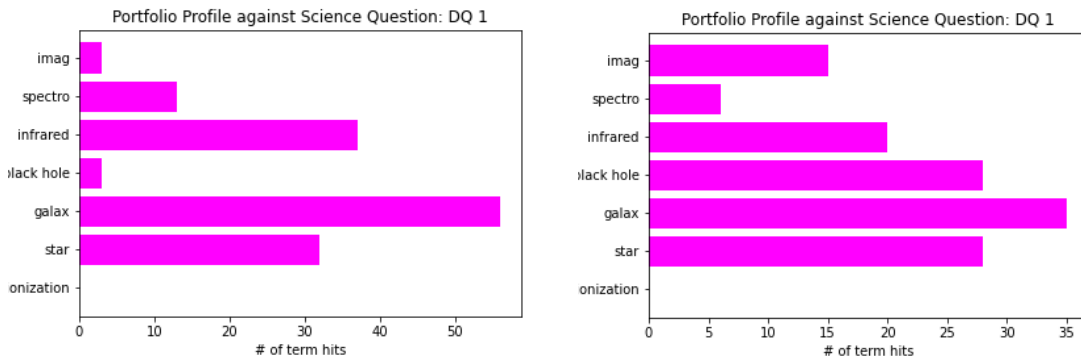


Figure 34: Topic/term distributions for science questions DQ-1 (under ‘Panel on Galaxies’).

Finally, we can compare both proposals directly and attain similar bar charts showing ‘gains’ and ‘losses’ in scientific impact across both the total relevance profiles and for specific science questions. For these portfolios, and for science question DQ-1, we receive the following comparison charts:

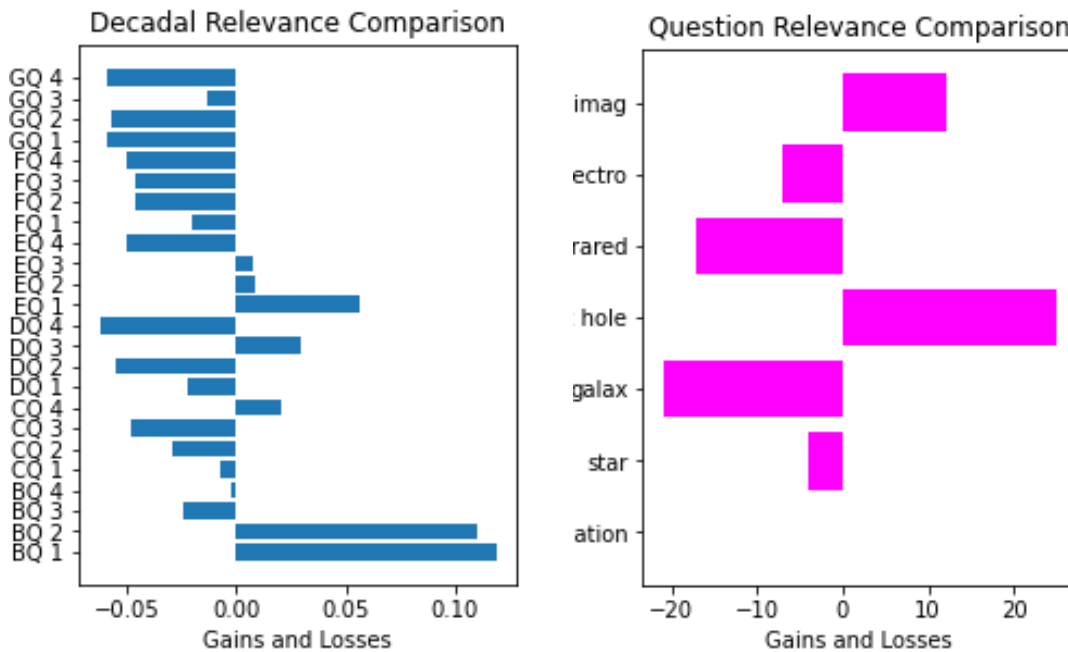


Figure 35: The comparison charts showing how portfolio 2's science impact compares with portfolio 1's scientific impact.

Discussion

Across all three mission concepts, AstroNLP provided a STG for each and produced their individual mission-level decadal relevance profile. What is also noteworthy is AstroNLP's ability to explore various portfolio decadal relevancies at the behest of a reviewer's choice selection of concepts.

When considering the STGs, due to performance implications, the implementation of POS and ontology enforcement heuristics is necessary to increase the readability and quality of the graphs. This is especially true when using higher epoch NER transformer models as a higher frequency of noisy entities will typically appear in

the output. Additionally, due to the forced filtering of entities with no relations attached to them, it is highly possible that much mission critical information is missed in any given STG. However, these graphs do already contain very relevant information that a reviewer would find useful when generating recommendations in the decadal survey. We also recognize that improving the quality of the STGs through additional training data or adding a various assortment of filtering rules (while both are incredibly necessary in this case) can be a Sisyphean effort given the nature of transformer-based models. Regardless, any future work on this endeavor should aim to substantially increase the annotation data pool by at least an order of magnitude so as to increase the fidelity of STG extraction.

From the perspective of decadal relevance analysis, we can very clearly visualize a mission/portfolio's relevance to specific science questions given in the decadal survey and also get a pretense in regards to how a sample portfolio compares with another. In the mission concepts examined, the outputs of the AstroNLP tool regarding a concept's most relevant science panel and questions show very plausible results given the descriptions contained in those concept proposals (e.g. LUVOIR's orientation towards the search and characterization of habitable exoplanets was well captured by AstroNLP). At the portfolio level, these outputs become even more intriguing as a reviewer can visually inspect and optimize its portfolio towards a desirable distribution of science topics (i.e. tailor to specific panels, or attempt to achieve a relatively even distribution across all science questions).

CHAPTER V

CONCLUSIONS

Computers in the modern era can support and enhance a wide variety of previously human-dominated tasks thanks, in part, to the rise of computer processing abilities and artificial intelligence. In this thesis, we covered one very specific area where natural language processing, a subtopic of artificial intelligence, can enhance what is still a human dominated activity (reviewing and adjudicating proposals).

Specifically, this work discussed AstroNLP; a tool capable of providing support to reviewers involved in evaluating space-based astronomy and astrophysics mission concepts. AstroNLP gives the reviewer the ability to upload several concept proposals and automatically determine 1) their science traceability and 2) their scientific relevance to a program. Discussions regarding the functions and processes powering AstroNLP were also provided, as well as mentions on current limitations still affecting AstroNLP.

To demonstrate the ability of AstroNLP, we provided a use case looking at three mission concepts submitted to the 2020 Astrophysics Decadal Survey; LUVOIR, OST, GEP. Among these three concepts, we portrayed their STGs extracted directly from their decadal proposals and showed their respective relevancies across the decadal's science panels and questions. We also showed the use-case of portfolio comparison detailing how two different portfolio's science relevancies can be analyzed directly against each other.

Whilst far from perfect, automated science traceability extraction and relevance assessment of mission concepts holds enormous implications not only for astrophysics, but for other scientific and non-scientific domains (e.g. defense). The tools and workflows discussed in this thesis can, in varying degrees, be adapted to these other domains to support similar processes across program formulation.

It is also worthy to mention, a-posteriori of this work, that it is vital when developing semantic technologies for a domain-specific use case to not only adapt the tools for said domain, but to understand the context of the given domain. As with many machine learning models, the performance of those works are only as good as the data provided during training and is therefore essential to understand the semantic structure, language habits, nuances, and ‘features’ of the domain material so as to given such a model the best chance at performing. This is also plays into the argument of verification for the work generated in this thesis, as model verification is paramount before any said implementation of this work can be initiated. We’ve provided a first step towards model verification through our provided results and performance metrics but recognize that a more robust verification framework will be necessary (e.g. comparing STGs with human-developed STMs from the same textual source) in future work.

REFERENCES

- [1] NASA. NASA Technology Roadmap. 2015.
<https://www.nasa.gov/offices/oct/home/roadmaps/index.html>. Accessed: January 20, 2022.
- [2] NASA. NASA Systems Engineering Handbook. NASA SP-2016-6105 Rev2. 2016.
- [3] Department of Defense. Mission Engineering Guide. *Washington, DC*. November, 2020.
- [4] Weiss, J., Smythe, D., Lu, W. Science traceability. *2005 IEEE Aerospace Conference*. 2005. pp. 292-299. doi: 10.1109/AERO.2005.1559323.
- [5] NASA. 2020 Decadal Survey Planning. 2022.
<https://science.nasa.gov/astrophysics/2020-decadal-survey-planning>. Accessed: January 20, 2022
- [6] Guariniello, C., Marsh, T., Porter, R., Crumbly, C., DeLaurentis, D. Artificial Intelligence Agents to Support Data Mining for SoS Modeling of Space Systems Design. In *2020 IEEE Aerospace Conference*. 2020. pp. 1-11. doi: 10.1109/AERO47225.2020.9172802.
- [7] Jones-Wilson, L., & Susca, S. (2017, March). A framework for extending the Science Traceability Matrix: application to the planned Europa mission. In *2017 IEEE Aerospace Conference* (pp. 1-14). IEEE.

- [8] Jones-Wilson, L., Susca, S., Reinholtz, K. Project-domain Science Traceability and Alignment Framework (P-STAF): Analysis of a payload architecture. In *2018 IEEE Aerospace Conference*. 2018. pp. 1-16. doi: 10.1109/AERO.2018.8396634.
- [9] National Research Council. Pathways to Discovery in Astronomy and Astrophysics for the 2020s. *Washington, DC: The National Academies Press*. 2021
- [10] Simpson, B., Selva, D., Richardson, D. Extracting Science Traceability Graphs from Mission Concept Documentation using Natural Language Processing. In *2022 AIAA: Science and Technology Forum*. 2022. doi: 10.2514/6.2022-1182
- [11] Thronson, H., Thomas, B., Barbier, L., Buonomo, A. Transforming Science and Technology Prioritization Processes Using Artificial Intelligence. *237th AAS Conference*. Poster 541.10. January 2021.
- [12] Manning, C., Schutze, H. 1999. Foundations of Statistical Natural Language Processing. *The MIT Press*. Cambridge, Mass.: MIT Press.
- [13] Salado, A. and Nilchiani, R. (2014), A Categorization Model of Requirements Based on Max-Neef's Model of Human Needs. *Syst. Engin.*, 17: 348-360.
<https://doi.org/10.1002/sys.21274>
- [14] Lian, X., Liu, W., Zhang, L. Assisting engineers extracting requirements on components from domain documents. *Information and Software Technology*, Volume 118. 2020.

- [15] Mokammel, F., Coatanéa, E., Coatanéa, J., Nenchev, V., Blanco, E., & Pietola, M. (2018). Automatic requirements extraction, analysis, and graph representation using an approach derived from computational linguistics. *Syst. Eng.*, 21, 555-575.
- [16] Arellano, A.G., Carney, E.M., & Austin, M.A. (2015). Natural Language Processing of Textual Requirements. *ICONS 2015*.
- [17] Arellano, A.G. (2018). Frameworks for Natural Language Processing of Textual Requirements.
- [18] NASA. (2020). “NASA Technology Taxonomy”.
<https://www.nasa.gov/offices/oct/taxonomy/index.html>. Accessed: January 20, 2020
- [19] Zaibert, L. (2016). The Theory and Practice of Ontology. London: *Palgrave Macmillan* UK :Imprint: Palgrave Macmillan.
- [20] Cox, A.P., Nebelecky, C.K., Rudnicki, R., Tagliaferri, W.A., Crassidis, J.L., & Smith, B. (2016). The Space Object Ontology. 2016 *19th International Conference on Information Fusion (FUSION)*, 146-153.
- [21] Rovetto, R.J., Kelso, T.S., & O’Neil, D.A. (2020). Orbital debris ontology, terminology, and knowledge modeling. *Journal of Space Safety Engineering*, 7, 451-458.
- [22] Rovetto, R.J. (2017). An ontology for satellite databases. *Earth Science Informatics*, 10, 417-427.

- [23] Hennig, C., Viehl, A., Kämpgen, B., & Eisenmann, H. (2016). Ontology-Based Design of Space Systems. *SEMWEB*.
- [24] Dori, D., & Sillitto, H. (2017). What is a System? An Ontological Framework. *Syst. Eng.*, 20, 207-219.
- [25] Thomsen, E., Smith, B. (2018). Ontology-based fusion of sensor data and natural language. *Applied Ontology* 13 (4):295-333.
- [26] Gómez, M., Preece, A.D., Johnson, M.P., Mel, G.D., Vasconcelos, W.W., Gibson, C., Bar-Noy, A., Borowiecki, K., Porta, T.F., Pizzocaro, D., Rowaihy, H., Pearson, G., & Pham, T. (2008). An Ontology-Centric Approach to Sensor-Mission Assignment. *EKAW*.
- [27] Deitz, P.H., Michaelis, J.R., Bray, B., & Kolodny, M.A. (2016). The Missions & Means Framework (MMF) Ontology : Matching Military Assets to Mission Objectives.
- [28] Ghaisas, S., & Ajmeri, N. (2013). Knowledge-assisted ontology-based requirements evolution. In *Managing requirements knowledge* (pp. 143-167). *Springer*, Berlin, Heidelberg.
- [29] Bernardi, T.L., Rabello, R.D., & Cervi, C.R. (2016). An Ontology-Based Approach to Use Requirements Engineering in Portals of Transparency. *ONTOBRAS*.
- [30] Castaneda, V., Ballejos, L.C., Caliusco, M.L., & Galli, M.R. (2010). The Use of Ontologies in Requirements Engineering. *Global Journal of Research In Engineering*, 10.

- [31] Siegemund, K., Thomas, E., Zhao, Y., Pan, J.Z., & Assmann, U. (2011). Towards Ontology-driven Requirements Engineering.
- [32] Natural Language Toolkit, NLTK Project, Ver. 3.6.7, <https://www.nltk.org/>
- [33] spaCy, Explosion, Ver. 3.0, <https://spacy.io/>
- [34] Thomas, A., & Sangeetha, S. (2019). An innovative hybrid approach for extracting named entities from unstructured text data. *Computational Intelligence*, 35, 799 - 826.
- [35] Al-Aswadi, F.N., Chan, H.Y., & Gan, K.H. (2019). Automatic ontology construction from text: a review from shallow to deep learning trend. *Artificial Intelligence Review*, 53, 3901 - 3928.
- [36] Berquand, A., Darm, P., & Riccardi, A. (2021). SpaceTransformers: language modeling for space systems. *IEEE Access*, 9, 133111-133122.
- [37] Fu, J., Liu, P., & Neubig, G. (2020). Interpretable multi-dataset evaluation for named entity recognition. *arXiv preprint arXiv:2011.06854*.
- [38] Patil, N., Patil, A., & Pawar, B. V. (2020). Named entity recognition using conditional random fields. *Procedia Computer Science*, 167, 1181-1188.
- [39] Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.
- [40] Krishnan, J., Coronado, P., Purohit, H., & Rangwala, H. (2020). Common-Knowledge Concept Recognition for SEVA. *ArXiv*, abs/2003.11687.

- [41] Crossland, T., Stenetorp, P., Riedel, S., Kawata, D., Kitching, T.D., & Croft, R. (2019). Towards Machine-assisted Meta-Studies: The Hubble Constant. *ArXiv*, abs/1902.00027.
- [42] Li, J., Sun, A., Han, J., & Li, C. (2020). A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*.
- [43] Yadav, V., & Bethard, S. (2019). A survey on recent advances in named entity recognition from deep learning models. *arXiv preprint arXiv:1910.11470*.
- [44] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- [45] Pleiss, G., Zhang, T., Elenberg, E. R., & Weinberger, K. Q. (2020). Identifying mislabeled data using the area under the margin ranking. *arXiv preprint arXiv:2001.10528*.
- [46] Pawar, S., Palshikar, G. K., & Bhattacharyya, P. (2017). Relation extraction: A survey. *arXiv preprint arXiv:1712.05191*.
- [47] Bach, N., & Badaskar, S. (2007). A review of relation extraction. *Literature review for Language and Statistics II*, 2, 1-15.
- [48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

- [49] Park, S., & Kim, H.M. (2020). Improving the Accuracy and Diversity of Feature Extraction From Online Reviews Using Keyword Embedding and Two Clustering Methods. *DAC 2020*.
- [50] Cvitanić, T., Lee, B., Song, H.I., Fu, K., & Rosen, D.W. (2016). LDA v. LSA: A Comparison of Two Computational Text Analysis Tools for the Functional Categorization of Patents. *ICCBR Workshops*.
- [51] Chinchor, N.A., & Sundheim, B.M. (1993). MUC-5 evaluation metrics. *MUC*.
- [52] Neuhaus, H., & Compton, M. (2009). The semantic sensor network ontology. In *AGILE workshop on challenges in geospatial data harmonisation*, Hannover, Germany (pp. 1-33).
- [53] PDF Parser. Jake Stockwin. Ver. 0.8.0. <https://py-pdf-parser.readthedocs.io/en/latest/index.html>.
- [54] Pdfminer.six. Yusuke Shinyama, Philippe Guglielmetti, & Pieter Marsman, Ver. 20201018, <https://pdfminersix.readthedocs.io/en/latest/>.
- [55] CoreNLP. Stanford NLP Group. Ver. 4.2.2. <https://stanfordnlp.github.io/CoreNLP/>.
- [56] OpenNLP. The Apache Software Foundation. Ver. 1.9.3. <https://opennlp.apache.org/download.html>.
- [57] UBIAI. UBIAI Web Services. <https://ubiai.tools/>. Accessed: January 20, 2022
- [58] NASA Space Science Data Coordinated Archive. NASA. <https://nssdca.gsfc.nasa.gov/>. Accessed: January 20, 2022

- [59] Astrophysics Data System. Smithsonian Astrophysical Observatory.
<https://ui.adsabs.harvard.edu/>. Accessed: January 20, 2022
- [60] GraphViz – Graph Visualization Software. GraphViz. Ver. 2.47.1.
<https://graphviz.org/download/>.
- [61] PyQt5. Riverbank Computing Limited. Ver. 5.15.6.
<https://pypi.org/project/PyQt5/>.

APPENDIX A

SCIENCE PANEL AND SCIENCE QUESTION TOPIC/TERM LISTS

	A	B	C	D
1	Panel	SQLabel	Science Questions	Area(s) of Unusual Discovery Potential
2	Compact Objects and Energetic			
3	Phenomena	BQ.1	What are the mass and spin distributions of neutron stars and stellar mass black holes?	Transforming our View of the Universe by
4		BQ.2	What powers the diversity of explosive phenomena across the electromagnetic spectrum?	
5		BQ.3	What do some compact objects eject material at nearly-light-speed jets, and what is that material made of?	
6		BQ.4	What seeds supermassive black holes and how do they grow?	
7	Cosmology	CQ.1	What set the hot Big Bang in motion?	The Dark Ages as a Cosmological Probe
8		CQ.2	What are the properties of dark matter and the dark sector?	
9		CQ.3	What physics drives the cosmic expansion and the large-scale evolution of the universe?	
10		CQ.4	How will measurements of gravitational waves reshape our cosmological view?	
11	Galaxies	DQ.1	How did the intergalactic medium and the first sources of radiation evolve from cosmic dawn through the epoch of reionization?	Mapping the Circumgalactic Medium and
12		DQ.2	How do gas, metals, and dust flow into, through, and out of galaxies?	
13		DQ.3	How do supermassive black holes form and how is their growth coupled to the evolution of their host galaxies?	
14		DQ.4	How do the histories of galaxies and their dark matter halos shape their observable properties?	
15	Exoplanets, Astrobiology, and the Solar System	EQ.1	What is the range of planetary system architectures, and is the configuration of the solar system common?	The Search for Life on Exoplanets
16		EQ.2	What are the properties of individual planets, and which processes lead to planetary diversity?	
17		EQ.3	How do habitable environments arise and evolve within the context of their planetary systems?	
18		EQ.4	How can signs of habitable life be identified and interpreted in the context of their planetary environments?	
19	Interstellar Medium and Star and Planet Formation	FQ.1	How do star-forming structures arise from, and interact with, the diffuse ISM?	Detecting and Characterizing Forming Planets
20		FQ.2	What regulates the structures and motions within molecular clouds?	
21		FQ.3	How does gas flow from parsec scales down to protostars and disks?	
22		FQ.4	Is planet formation fast or slow?	
23	Stars, the Sun, and Stellar Populations	GQ.1	What are the most extreme stars and stellar populations?	"Industrial Scale" Spectroscopy
24		GQ.2	How does multiplicity affect the way a star lives and dies?	
25		GQ.3	What would stars look like if we view them like we do the Sun?	
		GQ.4	How do the Sun and other stars create space weather?	

Figure 36: Panel and question portion of the knowledge based used as a reference guide for users of the AstroNLP tool so that relevance charts can be specifically pinpointed to specific scientific areas.

SQLLabel	Relevant Terms									
BQ 1	neutron star	black hole	timing	x-ray	imag					
BQ 2	neutron star	black hole	timing	x-ray	ultraviolet	imag	polarimet			
BQ 3	jet	composit	accelerat	optical	infrared	spectro	polarimet	imag		
BQ 4	black hole	redshift	optical	infrared	spectro	polarimet	imag			
CQ 1	gravitational wav	mapp	fluctuat							
CQ 2	dark matter	dark sector	timing	optical	infrared	ultraviolet	spectro	pulsar	interferomet	imag
CQ 3	neutrino	cosmic expansion	hubble constant	spectro	optical	infrared	cmb	interferomet		
CQ 4	gravitational wav	timing	dark age	cosmogr	imag	optical				
DQ 1	epoch of reionizat	star	galax	black hole	infrared	spectro	imag			
DQ 2	galax	gas	ultraviolet	x-ray	optical	infrared	spectro	metal	dust	
DQ 3	supermassive bla	time-domain sur	x-ray	ultraviolet	optical	infrared	spectro	imag		
DQ 4	milky way	galax	redshift	optical	infrared	spectro	x-ray			
EQ 1	planet	solar system	habitat	mass	radial velocit	timing	structure			
EQ 2	planet	atmosphere	imag	ultraviolet	infrared	mass	radial velocit	astrometry	spectro	
EQ 3	planet	habitable	star	imag	ultraviolet	infrared	optical	spectro	x-ray	
EQ 4	biosignature	planet	spectro	ultraviolet	infrared					
FQ 1	star	interstellar mediu	densit	temperature	x-ray	spectro	ultraviolet	optical	infrared	interferomet
FQ 2	molecular cloud	velocit	densit	structure	infrared	interferomet	spectro			
FQ 3	gas	protostar	formation	structure	optical	infrared	interferomet	spectro		
FQ 4	planet	formation	structure	atmosph	infrared	spectro	interferomet	optical		
GQ 1	star	optical	infrared	spectro	temperature	velocit	ultraviolet	radii	brown dwarf	interferomet
GQ 2	star	optical	infrared	ultraviolet	spectro	brown dwarf	temperature	interferomet		
GQ 3	sun	interferomet	optical	infrared	ultraviolet	x-ray	spectro	brown dwarf	spectropolarimet	
GQ 4	sun	spectropolarimet	optical	infrared	ultraviolet	spectro				

Figure 37: Term/topic map portion of knowledge base. Each row corresponds to a particular science question provided in the panel and question sheet of the excel document.