Annotation Guidelines for Astro-NER

1. Scientific Entity Definitions

- **AstrObject:** All concepts representing astronomical objects, e.g. black holes.
- **AstroPortion:** All concepts representing portions of astronomical objects which are not astronomical objects themselves, e.g. sunspots.
- **ChemicalSpecies:** Atomic elements such as element names from the periodic table, atoms, nuclei, dark matter, e.g. Fe.
- **Instrument:** Names of measurement instruments, including telescopes, e.g. Large Hadron Collider.
- **Measurement:** Measured observational parameters or properties (both property and value), e.g. frequency.
- **Method:** Abstractions which are commonly used to support the solution of the investigation, e.g. minimal supersymmetrical model.
    - ⇒ in case of overlap, prioritize Method over all other entity types except ResearchProblem
- **Morphology:** Geometry or morphology of astronomical objects or physical phenomena, e.g. asymmetrical.
- **PhysicalQuantity:** Properties of physical phenomena interacting, e.g. gravity.
- **Process:** Phenomenon or associated process, e.g. Higgs boson decay.
- **Project:** Survey or research mission, e.g. the dark energy survey.
- **ResearchProblem:** The theme of the investigation, e.g. final state hadronic interactions.
    - ⇒ in case of overlap, prioritize ResearchProblem over all other entity types
- **SpectralRegime:** Observed or analyzed electromagnetic spectrum, e.g. mega electron volt.

2. Scientific Entity Examples

AstrObject

Dilatonic	BTZ	black	holes	with	power	-	law	field
B-Morphology	B-AstrObject	I-AstrObject	I-AstrObject	O	B-PhysicalQuantity	I-PhysicalQuantity	I-PhysicalQuantity	I-PhysicalQuantity

AstroPortion

Chaos	in	charged	AdS	black	hole	extended	phase	space
O	O	B-Morphology	I-Morphology	B-AstroPortion	I-AstroPortion	I-AstroPortion	I-AstroPortion	I-AstroPortion

ChemicalSpecies

Electroexcitation	of	the	Δ+(1232	)	at	low	momentum	transfer
B-Process	O	O	B-ChemicalSpecies	I-ChemicalSpecies	O	B-PhysicalQuantity	I-PhysicalQuantity	I-PhysicalQuantity

Instrument

Electromagnetic	processes	in	ultra	-	peripheral	Pb+Pb	collisions	with	ATLAS
B-Process	I-Process	O	B-Process	I-Process	I-Process	I-Process	I-Process	O	B-Instrument

Measurement

Interference	of	dark	matter	solitons	and	galactic	offsets
O	O	B-ChemicalSpecies	I-ChemicalSpecies	I-ChemicalSpecies	O	B-Measurement	I-Measurement

Morphology

Light	sterile	neutrino	in	the	minimal	extended	seesaw
B-Morphology	B-Morphology	B-ChemicalSpecies	O	O	B-Method	I-Method	I-Method

PhysicalQuantity

Absence	of	log	correction	in	entropy	of	large	black	holes
B-ResearchProblem	I-ResearchProblem	I-ResearchProblem	I-ResearchProblem	O	B-PhysicalQuantity	O	B-Morphology	B-AstrObject	I-AstrObject

Process

Charged	-	lepton	mixing	and	lepton	flavor	violation
B-Process	I-Process	I-Process	I-Process	O	B-Process	I-Process	I-Process

Project

JUNO	:	A	Next	Generation	Reactor	Antineutrino	Experiment
B-Project	O	O	O	O	B-Instrument	B-ChemicalSpecies	O

SpectralRegime

Two	-	body	photodisintegration	of	3He	between	7	and	16	MeV
B-ResearchProblem	I-ResearchProblem	I-ResearchProblem	I-ResearchProblem	O	B-ChemicalSpecies	O	B-SpectralRegime	I-SpectralRegime	I-SpectralRegime	I-SpectralRegime

3. Annotation Scheme

3.1 Semantic Criteria

Definitions/descriptions of final semantic types

Principles:

- Give precedence to annotating “research problem” in cases where an entity is both a research problem and another entity type (e.g. method)

- Deciding a Research problem
    - Resolving ambiguities whether to split a phrase on the preposition “of” or not where if the phrase was split the first constituent would be the research problem and the second constituent phrase would be one among the other candidates but most likely a process or a method.
        - If the part of the phrase preceding “of” is too generic, then consider the whole unit including the prepositional phrase as a research problem.
        - Furthermore, if the part of the phrase following “of” is a process, then in most cases annotate the whole unit including the prepositional phrase as a research problem.
        - ~~If the part of the phrase following “of” is of type matter, then in most cases annotate the first part of the phrase as research problem and the second part as matter. This would hold unless the first part of the phrase is too generic. So the inclination is to try to favor annotating matter entities as far as possible.~~

- Err on the side of annotating more astronomy-specific entities rather than more general options (e.g. “atomic element” rather than “matter”, where possible)

3.2 Linguistic Criteria

There are no restrictions on the morphosyntactic form of terms. However, some principles apply:

- **Noun phrases** without articles are preferred wherever possible.

- **Verbs and verb phrases** are also allowed.

- **The most precise text reference possible**, including any modifiers and generic nouns, should be annotated as one unit.

  - “carbon atoms in graphene” and ***not*** “carbon atoms”, “graphene”
  - “sequential labeling approach” and ***not*** “sequential labeling”
  - For entity types like method and process, annotate individually all components which can stand on their own
  - For entity types like matter, location, and astronomical object, annotate the most specific span possible

3.3 Formal Criteria

- **Length:** terms may be one or more words long.

- **Span:** annotated terms should have an uninterrupted span, no linking across phrases.

- **Determiners:** do not include articles.

- **Term-abbreviation sequence:** a sequence like “machine translation (MT)” contains the same term twice, first the long form and then the abbreviation. This should be annotated as a single unit containing both forms.

- **Terms broken by abbreviations:** a sequence like “machine translation (MT) evaluation” contains the abbreviation of a general term (“MT”) inserted into the term of a more specific concept (“machine translation evaluation”). The entire sequence should be annotated as one unit.

- **Proper nouns:** some proper nouns may correspond to the semantic entity types defined above (e.g. WordNet as a *resource*), in which case they should be annotated.

- **Adjectival modifiers:** some terms are modified by adjectives. If removing the adjective changes the meaning of the term, then the adjective should be annotated with the noun it modifies as one lexical unit.
  - “Statistical machine translation” → annotate “statistical machine translation”, because there are other non-statistical approaches to machine translation and therefore statistical changes the meaning of the term from general to more specific
  - “Systematic pattern” → annotate only “pattern”, because patterns are by nature systematic and therefore removing “systematic” doesn’t change the meaning.

- **Conjunctions with ellipsis:** sometimes when two noun phrases both contain the same noun or noun phrase and are joined with a conjunction, the shared noun (phrase) may be stated explicitly only once, e.g. “supervised and unsupervised methods” means “supervised methods and unsupervised methods”. In this case the entire sequence should be annotated as one unit.

    - “machine and deep learning approaches” → annotate “machine and deep learning approaches” because “learning approaches” applies to both “machine” and “deep”
    - “TREC 2003 and TREC 2004” → annotate “TREC 2003” and “TREC 2004”

- **Prepositions:** terms can generally be split at prepositions, unless the preposition is in fact part of the term and modifies the term in an essential way
    - “automatic evaluation of machine translation and document summarization” → annotate “automatic evaluation” and “machine translation” and “document summarization” as each of these are individual terms
    - “part of speech tagging” → annotate “part of speech tagging”

- **Context:** annotate the most complete term corresponding the the *intended meaning* in the given context.
    - “This paper presents a maximum entropy word alignment algorithm for Arabic-English based on supervised training data.” → annotate “supervised training data” and ***not*** “supervised training” and “data” as two separate terms, because contextually the intended meaning is the single term “supervised training data”

- **Nested terms:** given an expression in which several concepts/terms are nested (e.g., “maximum entropy word alignment algorithm”), annotate the entire sequence as one term.

- **Incorrect spelling:** we assume that incorrect spellings will be rare in our dataset. If an incorrect spelling does occur, it should still be annotated.