# Labeling Guidelines for Ligand Selection in Nuclear Fuel Reprocessing

**Purpose**: This document provides comprehensive instructions for creating consistent and accurate labels for scientific literature used in ligand selection for nuclear fuel reprocessing. The goal is to ensure consistent and accurate labeling using Prodigy for training learning models in Named Entity Recognition (NER) and Information Extraction (IE) tasks. Customized categories will be used: Ligand Type, Chemical Stability, Thermodynamics, Kinetics, Loading Capacity, Operational Condition Range, Solubility, Dispersion Numbers, and Phase Disengagements.

**I General Labeling Workflow in Prodigy**

1. **Select the relevant text snippet**
   1. Highlight the portion of the text that matches a customized category.
   2. Avoid overlapping annotations; focus on distinct entities.
   3. Ensure the selected text is concise and contextually appropriate.
2. **Assign a category**
   1. Choose the most appropriate label from the following customized options:
      1. Ligand Type: Specific ligand names or abbreviations.
      2. Chemical Stability: Information on decomposition resistance or stability conditions.
      3. Thermodynamics: Details about energy properties, binding affinity, or enthalpy changes.
      4. Kinetics: Information on reaction rates or mechanisms.
      5. Loading Capacity: Maximum or optimal loading of material (e.g., metal ions, actinides).
      6. Operational Condition Range: Defined ranges for pH, temperature, or other environmental parameters.
      7. Solubility: Solubility characteristics in specific solvents or systems.
      8. Dispersion Numbers: Quantitative or qualitative metrics on phase dispersion.
      9. Phase Disengagements: Descriptions or data on phase separation behavior.
3. **Use Prodigy features for efficiency**
   1. Utilize keyboard shortcuts for fast navigation.
   2. Apply "accept/reject" options to maintain annotation quality.
   3. Use Prodigy’s pre-labeling feature to assist with repetitive labels.
4. **Review and export**
   1. Periodically review the annotations to ensure consistency.
   2. Export the labeled data in JSONL format for further processing.

**II Category-Specific Labeling Guidelines**

1. **Ligand Type**
   1. Definition: Specific ligand names or abbreviations.
   2. Guidance
      1. Use chemical nomenclature or abbreviations for ligand names (e.g., TODGA, T2EHDGA)
   3. Examples
      1. Highlight: "N,N,N′,N′-tetraoctyldiglycolamide"  
         Label: "Ligand Type”
      2. Highlight: "SO3-Ph-BTP”  
         Label: "Ligand Type"
      3. Highlight: "trialkyl phosphine oxide"  
         Label: "Ligand Type"
2. **Chemical Stability**
   1. Definition: The ligand's resistance to decomposition or activity under certain conditions.
   2. Guidance
      1. Highlight stability properties in specific environments (e.g., temperature, pH, or solvent).
      2. Avoid including unrelated operational or thermodynamic data.
   3. Examples:
      1. Highlight: "SO3-PhBTP remains active in nitric acid up to 3 M."  
         Label: "Chemical Stability"
      2. Highlight: "This ligand decomposes in alkaline solutions."  
         Label: "Chemical Stability"
3. **Thermodynamics**
   1. Definition: Energy-related properties, such as binding affinities or reaction heat.
   2. Guidance
      1. Focus on specific thermodynamic measures or behaviors.
      2. Avoid operational or process-related details unless they directly tie into thermodynamics.
   3. Examples
      1. Highlight: "The ligand shows high binding affinity for actinides."

Label: "Thermodynamics"

* + 1. Highlight: "Exothermic reaction observed during ligand interaction."

Label: "Thermodynamics"

1. **Operational Condition Range**
   1. Definition: The process conditions (e.g., temperature, pH, solvent) under which the ligand is effective.
   2. Guidance
      1. Include measurable parameters and avoid vague descriptions.
   3. Examples
      1. Highlight: "Operational range includes temperatures up to 70°C."

Label: "Operational Condition Range"

* + 1. Highlight: "Ligand works best in acidic environments."

Label: "Operational Condition Range"

1. **Kinetics**
   1. Definition: Reaction rates, mechanisms, or kinetic properties.
   2. Example
      1. Highlight: "Fast extraction within 10 seconds"
      2. Label: “Kinetics”
2. **Loading Capacity**
   1. Definition: Maximum or optimal material loading.
   2. Example
      1. Highlight: "Can load up to 0.5 g/L"
      2. Label: “Loading Capacity”
3. **Solubility**
   1. Definition: Solubility characteristics in solvents or systems.
   2. Example
      1. Highlight: "Highly soluble in dodecane"
      2. Label: “Solubility”
4. **Dispersion Numbers**
   1. Definition: Metrics on phase dispersion efficiency or behavior.
   2. Example
      1. Highlight: "Dispersion number of 0.25 achieved"
      2. Label: “Dispersion Numbers”
5. **Phase Disengagements**
   1. Definition: Descriptions or data on phase separation performance.
   2. Example
      1. Highlight: "Rapid phase separation within 5 seconds"
      2. Label: “Phase Disengagements”

**III Workflow for Labeling**

1. **Read and Understand the Snippet:**
   1. Identify the key information related to ligand selection.
2. **Select the Relevant Category:**
   1. Assign the label to the most appropriate category using these guidelines.
3. **Review for Consistency:**
   1. Ensure that the label adheres to the word count and format rules.

**III Workflow Adjustments**

1. **Annotation Tool Preparation**
   1. Update Prodigy’s configuration to include the customized categories:

A screenshot of a computer program

Description automatically generated

1. **Text Selection**
   1. Use Prodigy’s highlighting tools to select text that matches one of the categories.
   2. Avoid overlapping or ambiguous highlights to maintain clarity.
2. **Quality Control**
   1. Use Prodigy’s review feature for double annotation by two reviewers to ensure label accuracy.
   2. Resolve discrepancies collaboratively or with predefined rules.
3. **Data Export**
   1. Export annotated data in JSONL format for model training.
   2. Include metadata such as document source or section reference for context.

**IV Additional Notes**

* Ambiguity: If a snippet fits multiple categories, create labels for all relevant categories.
* Consistency: Regularly review and update labels for uniformity across the dataset.
* Collaboration: Confirm labels with the team during the initial phase to ensure alignment.

**V Enhancements for NER and IE Compatibility**

1. **Entity Boundaries:**
   1. Clearly define the boundaries of entities within text snippets.
   2. Guidance:
      1. Highlight only the most relevant words for the label to help NER models detect precise entities.
      2. Avoid including extraneous text or surrounding context that might confuse the model.
   3. Example:
      1. Snippet: "SO3-PhBTP remains active in nitric acid up to 3 M."
      2. Label: "SO3-PhBTP" (Entity for Ligand Type).
      3. Label: "3 M Nitric Acid" (Entity for Operational Condition).
2. **Avoid Overlapping or Nested Labels:**
   1. Each entity should belong to a single category.
   2. If a snippet describes multiple attributes, split the text and create separate labels for each.
   3. Example:
      1. Snippet: "SO3-PhBTP remains active in 3 M nitric acid and separates actinides efficiently."
      2. Labels:
         1. Chemical Stability: "Active in 3 M Nitric Acid"
         2. Extraction Efficiency: "Efficient Actinide Separation"
3. **Attribute Value Extraction**
   1. For numerical or measurable data (e.g., temperature, pH, concentration), ensure that the label extracts the specific value or range clearly.
   2. Guidance:
      1. Include units (e.g., °C, M) where applicable.
      2. Avoid ambiguous phrases like "high" or "low" unless the text snippet provides comparative data.
   3. Example:
      1. Snippet: "The ligand is effective up to 70°C."
      2. Label: "Effective up to 70°C" (Operational Condition).
4. **Consistent Annotation Schema**
   1. Use a consistent format for all labels to make them compatible with NER training datasets (e.g., BIO tagging).
   2. Define tags based on categories:
      1. B-<Category>: Beginning of the entity.
      2. I-<Category>: Inside the entity.
      3. O: Outside the entity.
   3. Example (BIO format):
      1. Snippet: "SO3-PhBTP remains active in nitric acid up to 3 M."
      2. Tags:
         1. SO3-PhBTP → B-Ligand\_Type
         2. remains active in → O
         3. nitric acid → B-Operational\_Condition\_Range
         4. up to → O
         5. 3 M → B-Chemical\_Stability
5. **Data Preprocessing for NLP Tasks:**
   1. Ensure the labeled data is preprocessed for NER and IE tasks:
      1. Remove unnecessary punctuation and special characters.
      2. Tokenize text properly to match NER model input formats.

**VI Recommendations for Repeated Labels**

#### **Within the Same Article**

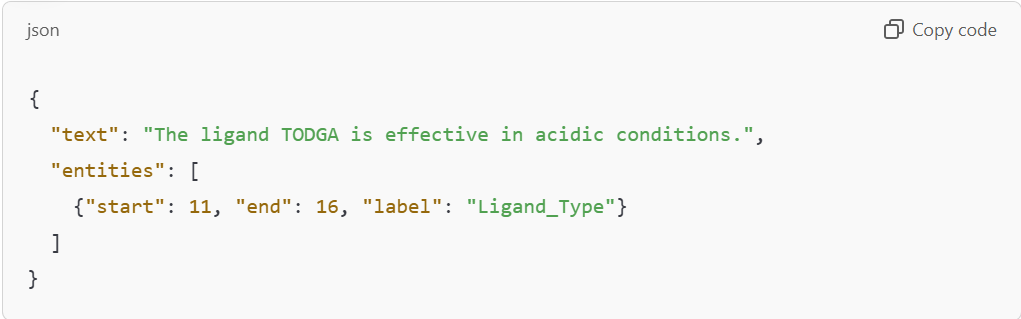
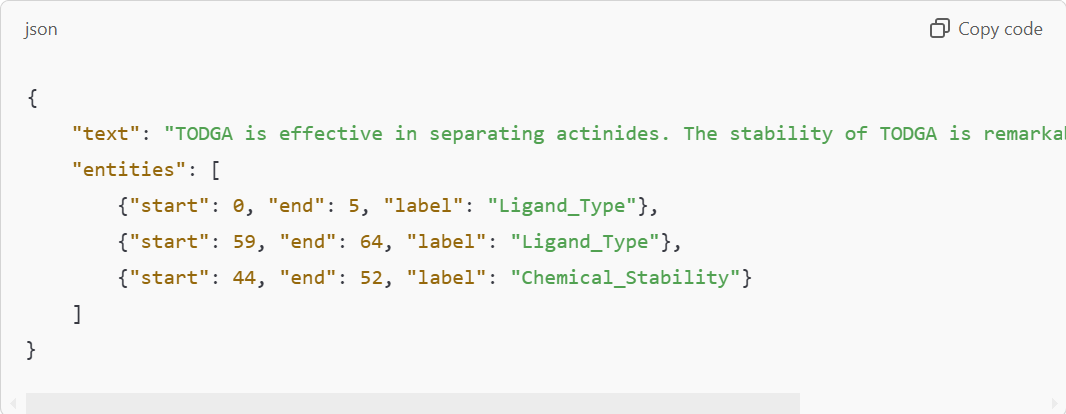
* 1. **Why Repeat Labels?**
     1. Each occurrence of an entity (e.g., "TODGA") might appear in a different **context** or describe a different attribute (e.g., stability, thermodynamics, or process compatibility). Labeling each instance helps the model learn these contextual variations.
  2. **When to Repeat Labels**:
     1. If the entity appears multiple times in unique **contexts**.
        1. Example:
           1. Snippet 1: "TODGA binds strongly with actinides."  
              Label: "TODGA" (Ligand Type).
           2. Snippet 2: "The stability of TODGA in nitric acid is remarkable."  
              Label: "TODGA" (Ligand Type) and "Stable in Nitric Acid" (Chemical Stability).
     2. If the **phrasing changes** even slightly, annotate it again to help the model learn the variation:
        1. Example:
           1. "TODGA shows high efficiency in separating actinides."
           2. "Actinide separation efficiency of TODGA is well-documented."
  3. **When NOT to Repeat**:
     1. If the exact same phrase is duplicated (verbatim) within the article, label it only once to avoid redundancy.
        1. Example:
           1. "TODGA is effective in acidic conditions." (appears twice, identical phrasing)
           2. Label it only the first time.

1. **Across Different Articles:**
   1. **Why Repeat Labels Across Articles?**
      1. Each article provides unique context, phrasing, and additional details that enhance the model’s understanding of entities and their attributes.
      2. Consistency ensures no information is missed.
   2. **Example**:
      1. 1: "TODGA has been widely used for actinide separation."
         1. Label: "TODGA" (Ligand Type) and "Actinide Separation" (Extraction Efficiency).
      2. Article 2: "The use of TODGA in acidic systems provides stability."
         1. Label: "TODGA" (Ligand Type) and "Stability in Acidic Systems" (Chemical Stability).
2. **For Large-Scale Automation**
   1. Once you have a significant number of examples for a repeated label (e.g., "TODGA"), you can:
      1. Use the model's predictions to annotate future occurrences automatically.
      2. Manually review the predictions for quality assurance.
3. **How Repeated Labels Benefit NER and IE Models**
4. **NER Model Training**:
   1. Repeated labels increase the dataset size, allowing the model to learn:
      1. Variations in how entities are described.
      2. The contexts in which entities occur.
   2. Example: A model exposed to "TODGA" in stability, thermodynamics, and extraction contexts will better generalize to unseen examples.
5. **IE Model Applications**
   1. Repeated labels ensure that all mentions of a critical entity are captured and linked during the extraction phase.
   2. This is especially important when aggregating information for downstream analysis (e.g., comparing ligand efficiencies across articles).

**VII Key Considerations**

* **Avoid Annotation Fatigue**
  + While repeated labeling is essential, balance the workload by prioritizing unique entities or contexts in initial experiments.
* **Leverage Automation**
  + Use tools like SpaCy or Hugging Face models to pre-label repeated entities after the initial manual annotations are complete.

**VIII How the Model Associates Labels with Text in the Document**

1. **Using Start and End Character Indexes**
   1. When labeling text for NER/IE, each label is tied to the exact position of the entity within the document.
   2. These positions are represented by the start and end character indexes of the labeled text.
   3. Example:
      1. Text Snippet: "The ligand TODGA is effective in acidic conditions."
      2. Label: "TODGA" as Ligand\_Type
      3. Output Format: 
   4. The model learns that the text between character 11 and 16 refers to "TODGA" as a "Ligand\_Type."
2. **Token-Based Annotation**
   1. If your text is tokenized (split into words or phrases), the label can also be tied to specific tokens.
   2. Example:
      1. Tokenized Text: ["The", "ligand", "TODGA", "is", "effective", "in", "acidic", "conditions."]
      2. Label: ["O", "O", "B-Ligand\_Type", "O", "O", "O", "O", "O"]
      3. Here:
         1. "B-Ligand\_Type" marks the start of the entity "TODGA."
         2. "O" marks tokens outside any labeled entity.
3. **Full Document Annotations**
   1. When labeling entire documents
      1. Labels are tied to text spans (subsections of text), not the entire document.
      2. Each entity within the document is annotated with its position and category
   2. Example: 
      1. Here, "TODGA" is labeled twice, but each instance is tied to its specific location.