A close-up of a document

Description automatically generatedA close-up of several papers

Description automatically generated

Data Files Needing Human Evaluation

Original Documents

Read ME -

**General Overview**

There are multiple CSV files that capture different layers of analysis on patent data, including:

1. **Chemical compound annotations** (e.g., whether a substance is an active ingredient, structural component, etc.).
2. **Patent Claims Analysis and chemical name extractions** (e.g., identifying compounds to chemicals , my result show 100 % accuracy for that data set - EP2266989B1 patent claim which is part of another study.<https://data.mendeley.com/datasets/6hykykmn65/1#:~:text=to%20extract%20chemical%20entities%20from,a%20patent%20corpus%20with%20annotations>
3. **Triplet extractions** from textual content (Subject–Predicate–Object relationships).

Although each file may vary slightly, they generally share a similar column structure with a few custom columns depending on the file’s purpose. Below is a description of all the main columns found across these datasets.

**Note**: Some columns appear only in specific files, but the explanation below covers all columns that may appear in any of the CSV files.

**Columns and Their Descriptions**

**1. node\_1**

**Meaning**: The first entity identified in a relationship or annotation. **Example**: "compound 1" or "L’OREAL"

**How to Fill**:

* This is typically provided by the automated or manual extraction process.
* Do **not** alter unless you see an obvious spelling mistake or mismatch.

**2. node\_2**

**Meaning**: The second entity identified in a relationship or annotation. **Example**: "hasFormula" could link node\_1 = "compound 1" with node\_2 = "(3,5-dibromo-4-hydroxy-phenyl)..." or for a patent relationship, node\_1 = "WO 2023/060387" with node\_2 = "WIPO".

**How to Fill**:

* Also typically provided by extraction processes.
* Update only if there is a clear error.

**3. edge**

**Meaning**: The type of relationship or link between node\_1 and node\_2. **Examples**: "related to", "is\_part\_of\_composition\_by\_weight", "hasFormula", or "is located at".

**How to Fill**:

* This is usually determined by the system’s extraction (or by the domain expert if manually assigned).
* Leave as-is unless an obvious mislabeling is found.

**4. chunk\_id**

**Meaning**: An internal identifier linking to a specific chunk of text from which the relationship or annotation was extracted. **Example**: "99e23d8937b44c9ba5218ea18f54ccab"

**How to Fill**:

* Typically auto-generated.
* No need to alter unless you know the chunk references are incorrect.

**5. count**

**Meaning**: A numerical counter for how many times a particular relationship or entity has been identified or appears. **Example**: 4 meaning it was found four times in the data.

**How to Fill**:

* Often automatically assigned by the extraction tool.
* Update only if you see a discrepancy in how many times it truly appears.

**6. node\_1\_attributes / node\_2\_attributes**

**Meaning**: JSON-like or dictionary-style entries providing metadata about each node. This can include:

* Whether the node is an active ingredient ("is\_active\_ingredient": True/False)
* Explanation or justification ("explanation": "why we consider it an active ingredient")
* Category of the entity (e.g., "chemical\_compound", "organisation", "concept", "functional\_group", etc.)
* Weight or percentage for the composition (if relevant, e.g., "active\_ingredient\_weight": "0.01% to 3%")

**How to Fill**:

* The content here is largely system-generated or domain-expert provided.
* Only correct or refine if there is an error or new domain insight that changes the classification (e.g., you discover that an ingredient is actually not active).

**7. Expert Answer**

**Meaning**: This is an **empty/optional** column intended for a human expert’s confirmation, correction, or final judgment about the extracted relationship or entity. It is where you note if the system’s extraction or label is correct, incorrect, or needs refinement.

**How to Fill**:

* **If the system’s extraction is correct**, you might write "Confirmed" or "Correct".
* **If the system’s extraction is incorrect**, you might write "Incorrect" or a more precise status such as "Needs Review".
* **If you need to clarify or specify** (e.g., the node is actually a different category), you can note that here.

**8. Short Justification**

**Meaning**: A concise textual explanation justifying your verdict in the *Expert Answer* column.

**How to Fill**:

* Provide **one or two sentences** explaining **why** you consider the extraction correct or incorrect.
* Examples:
  + "The chemical formula matches the described structure, so the classification is correct."
  + "This compound is misclassified; it’s actually a structural component, not an active ingredient."
* Keep it short and clear.

**9. ChunkContent (in the Simple Triplet Analysis file)**

**Meaning**: The raw text snippet from which the triplets (Subject, Predicate, Object) are extracted.

**How to Fill**:

* Usually left as provided. This is the actual snippet.
* Do not modify unless you need to correct typographical errors.

**10. Subject, Predicate, Object**

**Meaning**: The three parts of a subject–predicate–object triple extracted from the ChunkContent. **Examples**:

* Subject = "WO 2023/060387 A1"
* Predicate = "has publication date"
* Object = "20-Apr-23"

**How to Fill**:

* Typically done by the extraction process.
* You only correct if the triple is obviously mislabeled or incomplete.

**11. Confidence**

**Meaning**: A numeric score (0 to 1 or 0% to 100%) indicating how certain the extraction system is about that Subject–Predicate–Object triple.

**How to Fill**:

* Automatically assigned by the tool.
* No changes unless you recalculate or have a reason to override it.

**How to Use the *Expert Answer* and *Short Justification* Columns**

1. **Review each row**: Look at the extracted relationship or annotation (e.g., “Compound X is an active ingredient with 0.5% weight”).
2. **Check correctness**: Decide whether the relationship or label is correct, partially correct, or entirely incorrect.
3. **Fill *Expert Answer***:
   * For correct entries, you can simply write “Correct” or “Yes” (some prefer “Verified”).
   * For entries that need changes, write something like “Incorrect – category should be structural component” or “Partial – missing concentration details.”
4. **Give a *Short Justification***: Provide a succinct reason for your answer:
   * E.g., “Confirmed correct because the patent text clearly states 0.5% by weight as an active ingredient” or “Incorrect classification; the text describes it as a polymeric excipient, not an active ingredient.”

**Goal**: The combination of these two columns forms a light human-curated “silver standard” or “gold standard” dataset, capturing both the machine’s output and the expert’s final decision with a short rationale.