Biology Curation Engagement

Standigm, Inc. (STANDIGM) & Excelra Knowledge Solutions Pvt., Ltd (EXCELRA), jointly referred to as 'Parties'

About EXCELRA

EXCELRA is a leading Informatics services company that leverages its extensive scientific knowledge base, technology and relevant domain expertise to provide intelligent data and analytics solutions to clients in the pharma and biotech industry. EXCELRA's expertise includes biological & pharmacological data extraction, indexing, abstraction and analysis. Excelra has a considerable experience in data structuring across the drug development value chain, right from Discovery, Translational, Clinical and RWE data. It has built infrastructure over the last 18 years that supports this endeavor to provide data structuring services to its pharma and biotech partners. This includes a workforce of 500+ in-house resources with diverse domain expertise, SMEs with PHD & PostDoc level experience and a technology team to support all IT related requirements to help manage the data. Additionally, Excelra is also compliant by ISO standards -QMS, BCMS, ISMS, GDPR, HIPAA.

About STANDIGM

STANDIGM applies cutting-edge AI technologies to drug discovery and development. Both AI and biology experts team up to build real-world AI models to enhance and accelerate drug discovery productivity. The STANDIGM AIs reveal the drug patterns and generate potential candidates. STANDIGM is well known for its AI platforms such as STANDIGM BEST, an AI-based leading material optimization platform, and STANDIGM InsightTM, a new AI-based indication and mechanism of action prediction platform.

Objective

STANDIGM would like to engage EXCELRA's Data extraction & expert manual validation services to help validate & annotate the following:

- ***Objective 1:*** Labeling gene-disease interactions present in biomedical literature abstracts provided as output from the client's AI platform.
- ***Objective 2:*** Labeling relations between biological entities in an abstract

These annotations will then be used by STANDIGM to train a machine learning model to identify and classify target-disease interactions as well as relationships between biological entities.

Workflow

The two objectives will be met in parallel. The workflow for each is as follows:

- ***Objective 1: Labeling gene-disease interactions present in biomedical literature abstracts***
    - *Input:* An excel sheet with a list of 2000 abstracts with predicted gene-disease associations labelled.
    - *Process:* For each record in the excel sheet genes and disease names present in the abstract shall first be identified. They shall then be labelled as "Associated" or "Not Associated". In addition to the label, the text from the abstract that establish the relationship shall also be captured.
    - *Output:* Manually validated gene to disease relationship for each abstract in an edited excel sheet.

- ***Objective 2: Labeling relations between biological entities in an abstract***
    - *Input:* JSON files for 2000 abstracts with predicted biological entities along with predicted relationships labeled.

- *Process:* For each record the input file shall be uploaded into the viewer available at http://textae.pubannotation.org/editor.html?mode=edit and validate the annotations in the following steps (details of category definitions available at https://sites.google.com/view/bionlp-ost19-agac-track/description:
  - Identify the trigger words in PubMed abstracts and annotated them as correct trigger labels or entities (Var, MPA, Interaction, Pathway, CPA, Reg, PosReg, NegReg, Disease, Gene, Protein, Enzyme).
  - Identify the themetic roles (ThemeOf, CauseOf) between trigger words.
  - Extract the gene-function change-disease link. There are 4 different kinds of function change that link gene and disease: Loss of Function(LOF), Gain of Function(GOF), Regulation(REG), Complex(COM). LOF and GOF means loss or gain of function, while REG means the neutral or unknown link, and COM means the function changes between the gene and disease are in more complex way that can hardly to determine whether they are LOF or GOF.

  In case there are additional trigger words, themetic roles and gene-function change-disease link identified, they shall also be added and appropriately labelled.
- *Output:* The edited JSON file shall then be downloaded for each abstract and submitted as output.

- **Pilot Phase:**
  Both objectives shall be subjected to an initial pilot phase for 100 records each which shall be annotated, and the result produced for STANDIGM's perusal to understand the following:
  - Productivity on an hourly basis for benchmarking the rest of the activity
  - Quality measures to ensure accuracy in data capture
  - A review of the guidelines and any possible changes to it.
  - A list of anomalies found, if any, in the annotation of the 100 abstracts

  Further, upon the receipt of such results produced for STANDIGM's perusal and understanding, STANDIGM may provide suggestions, comments or other feedback ("Feedback") to EXCELRA regarding such results and the Parties shall discuss in good faith on how to implement such Feedback. In the event EXCELRA fails to implement the Feedback discussed in good faith between the Parties or otherwise fails to produce the results to the standard required by STANDIGM within 15 working days of STANDIGM's receipt of the results, STANDIGM shall be entitled to terminate the current engagement.

- **Final Phase:**
  The remaining 1900 abstracts will be annotated in adherence to the benchmarks, guidelines and rules set in the pilot phase. EXCELRA shall implement any modification or improvement requested by STANDGIM within 15 working days from the completion of the Final Phase.

  Excelra shall provide outputs and deliverables in the prescribed format:
  - **Objective 1:** Excel sheet with manually validated gene-disease associations label
  - **Objective 2:** JSON files with manually validated relationship labels for biological entities

Logistics

- **Scientific Meetings:** EXCELRA shall conduct project update meetings with the counterpart team internal at STANDIGM to share the progress made as required/depending on the needs of partner/project deliverables (Teleconference/F2F, based on STANDIGM's availability).
- **Intellectual Property Rights:** STANDIGM will have the right on all the inventions or discoveries generated.
- **Resources:** EXCELRA has the right set of resources to support data extraction and structuring activities for STANDIGM's R&D initiatives. As part of the engagement, EXCELRA will also utilize its internal tools, technologies, databases and assign a dedicated Project and Alliance manager that will be the point of contact for the project.
- **Resources:** EXCELRA has the right set of resources to this activity- Data Curators, Data Wranglers and SME's. As part of the engagement, EXCELRA will also utilize its internal tools, technologies and assign a dedicated Scientific Project Coordinator and Alliance manager that will be the point of contact for the project.

Timelines

A team of ***6 curators*** shall be put to the task.

Daily Team Output: ~32 abstracts

Weekly Output: 160 full text articles

No. of weeks for 2000 abstracts: 13 weeks (3.25 months)

---------- End of Proposal ----------