This Innovation Note was written by
KNIME community member Fernando López Bello

Medical record mapping

Giving practitioners access to the information they need

It's essential that medical geneticists know what the latest updates and findings are within their domain. This gives physicians security in their decision making because they know they are supported by the latest knowledge developments in that area, for which many relevant publications are made available each month. It also means that state of the art assertions can be made – resulting in correct diagnoses and better treatment and analysis.

Here a framework (Fig. 1) is built to help physicians make informed decisions and recommendations. This is done by constructing an information integration based on relevant entities and relationships from biomedical literature and then mapping these to an actual medical record. At the start of the process, main literature sources (1, 2) are queried for ingestion of relevant abstracts. Since abstracts tend to contain short, assertive sentences, it's expected they provide conclusions with concrete cause-effects mentioned, and, in general, neither negative or ambiguous implications. Then an NLP pipeline (3) is performed on that literature, where entities are recognized and relations are inferred. These outputs create a mini ontology (4), a semantic representation of linked concepts. At the same time, medical records are ingested (5) and relevant concepts identified (6). The final result is a graph database where medical records are linked to literature assertions.

This solution is built from abstracts related to genomics dating from 1990 to today and resulting in nearly four million abstracts. All are taken from PubMed and parsed with KNIME nodes (see Fig 4.). Several Named Entity Recognition techniques are combined, including dictionary-based tagging and the Abner tagger. All are available directly within KNIME.
For relationship extraction, a set of generic pattern-based rules (examples in Table 1) that can be configured by the physician is defined. These rules express a relation as an association between entities via stem words, which can be processed with KNIME NLP nodes. To validate inferred relations, a Distant Supervision approach is used based on OMIM´s (Online Mendelian Inheritance in Man) ground truth. This technique requires a classifier where, among others, several strategies are leveraged based on K-nearest neighbors, Support Vector Machines, and Multi-Layer Perceptrons.

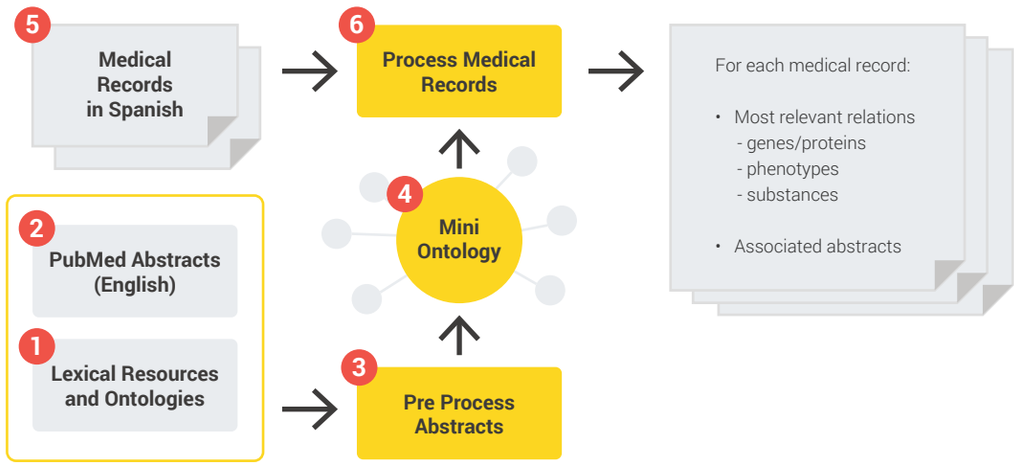

Fig. 1. Framework used to create graph databases where medical records are linked to literature assertions.

This Innovation Note was written by
KNIME community member Fernando López Bello

Finally, the mini ontology is created by combining a simplified version of Snomed CT ontologies - in both English and Spanish - with the recognized entities and relations from PubMed. These can be linked to entities in medical records written in Spanish. All Snomed CT concepts are loaded, as well as relationships between them and their translations. The mini ontology is stored as a graph in Neo4J, which can be queried on demand. The entire framework is depicted in Fig. 2. KNIME Analytics Platform runs workflows for ingesting and processing abstracts and medical records, as well as integrating the results into the mini ontology graph database. The final products can be consumed either by accessing the mini ontology and performing custom queries that apply to graph databases, or by searching literature on enriched text and inferring new relations.

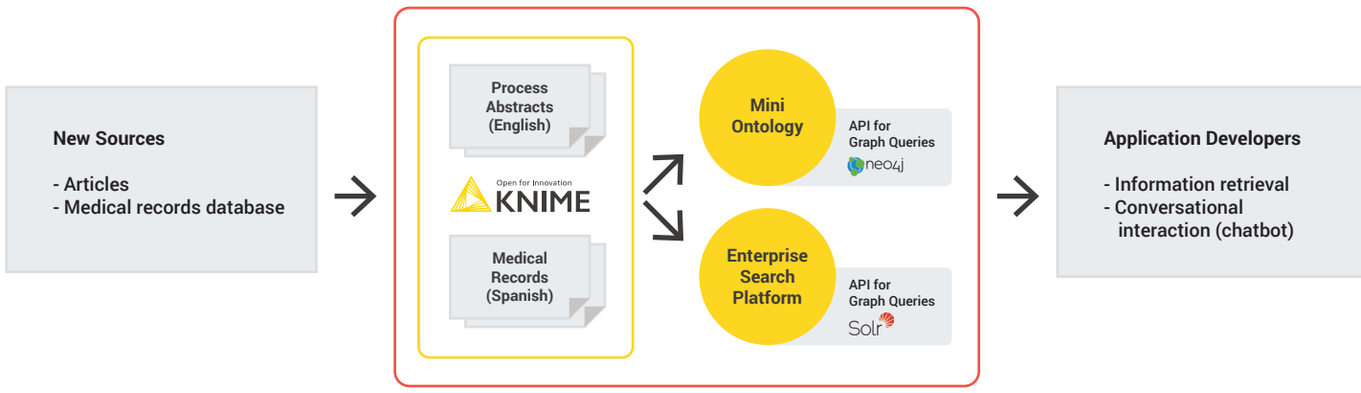

Fig. 2. An overview of the application framework.

Relationship Type	Entity Type 1	Cue Words	Entity Type 2
ASSOC	PROTEIN | GENE-SYMBOL	associated	DISEASE
CAUSE	PROTEIN | GENE-SYMBOL | PHENOTYPE	associated | lead to | leading to | responsabile | involved	DISEASE

Table 1. Examples of sets of generic pattern-based rules.

Other interesting features:

- Integration with FreeLing NLP tools (Universitat Politecnica de Catalunya, Spain), via sockets
- Node4J batch integration – mass export of entities and relations for the ontology
- Ready for enterprise search integration (Solr, ElasticSearch)
- Custom Java class for KNIME detailed timed log generation

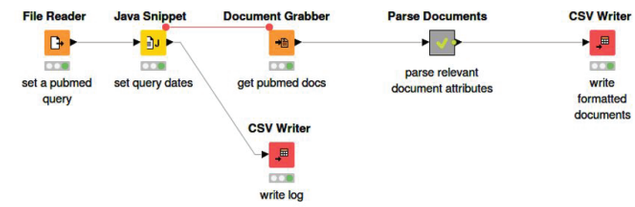

Fig. 4. Document acquisition workflow in KNIME Analytics Platform

Results:

From this project, the most significant results are:

- The mini ontology graph, which can be leveraged by external applications. This is an ongoing effort, as the KNIME workflow is designed to incorporate new sources.
- Inferring new relations, for example gene X associated with disease Y, which were later verified as valid in newer OMIM releases.

The open source KNIME Analytics Platform made this project successful because all tasks could be done within one visual environment. It was easy to read in the different data sources and the machine learning functionality such as the classification nodes enabled non-coders and programmers to work independently and efficiently. In addition, the KNIME Textprocessing Extension provided all the functionality needed for natural language programming nodes such as POS Tagger and Entity Recognition.

Try it out for yourself!

This workflow is available on the KNIME Hub:
https://tinyurl.com/article-grabber-pubmed