Minutes Limagrain use case discussion

29-09-2015

**Present:** Nicolas Heslot (Limagrain), Anneke Sijbers (ODEX), Wytze Vlietstra (ODEX, minutes)

**Goal of meeting:** Obtain detailed use case description of Limagrain for the ODEX project. Create schematic representation of the workflow. Obtain pre-requirements for implementation of use case.

**General remarks:**

* During our discussion two use cases were described. Both the original, and an updated and now preferred use case have been discussed.
  + Use case 1: Connected maize traits to genes through QTL’s. (original use case)
  + Use case 2: Compare text mined claims in maize patents to text mined claims of earlier patents and other (scientific) literature. Create a citation network for patents to identify potential prior art. (updated use case)
* In the spirit of the exploratory nature of ODEX, we will first investigate the potential for the updated use case, although its required text mining component might have severe pitfalls.

**Notes:**

* During the meeting, Nicolas indicated that Limagrain prefers to deviate from their originally described use case, which connects maize traits to genes, to a use case more focused on patents.
* The goal of this use case would be the discovery of prior art for the claims sections of maize patents, either from other patents or from (scientific) literature.
* This would be achieved by comparing the text mined triples from patents to other patents and scientific literature.
* A citation network would be created for the patents going ~20 years back. Nicolas is willing to provide the patent set to be text mined (pdf), which should range into the hundreds.
* One important pitfall is the terminology to be used for the text mining. While biomedical terminology is well developed, plant terminology is likely to be less so.
* Nicolas has indicated he will investigate existing plant terminology resources (plant ontology/gramene?) and, if required, manually extend them with e.g. article indexing terms to cover all relevant terms for maize.
* Importantly, Nicolas has indicated Limagrain is willing to spend manhours to validate the text mining results and to extend/improve the terminology iteratively.
* It would be highly convenient if first text mining results are available before December, as there will be a meeting of IP experts.

**Pre-requisites for algorithm:**

* Successfully text mined (minimally) the claims section of the patent.
* Created a citation network for the subset of the patents.

**Ultimate output:**

1. Citation network of patents and literature with certain metadata (patent granted, company behind patent, etc.)
2. Visualized graph in which the overlapping triples between two or more patents can be shown.

**Potential expansions:**

* Limit included patents to those of the most important competitors
* Include actual patented sequences, so these can easily be Blasted.
* Claims sections often have a well cited expansion, created by the patent office. In this expansion the patent office reports on its own search for prior art. This could possibly be a valuable source of additional information.

****

Figure 1: Schematic representation of the citation network and its relevant components.

**Original use case:**

* The goal of the original use case was to mine patents to connect plant (maize) traits (e.g. drought resistance, yield, disease resistance) to specific genes.
* The preferred data source for this would also be patents, in addition to literature.
* In addition, orthologues from other species would be included (Arabidopsis).
* Reference genomes are publically available in Panzea DB and MaizeGDB.

**Future communication:**

* Wytze will serve as point of contact for Nicolas within ODEX, due to his close vicinity to the text mining component of ODEX.
* Nicolas will be both the technical and content point of contact within Limagrain
* Both have expressed a desire to move fast on this use case.