Integrating PubTator in the curation pipeline of SourceData

Lou Götz[1], Nancy George[2], Sara El-Gebali[2], Anastasia Chasapi[1], Isaac Crespo[1], Ioannis Xenarios[1], Robin Liechti[1] and Thomas Lemberger[2]

[1]: *Vital-IT, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland*
[2]: *SourceData, EMBO, Heidelberg, Germany*

In scientific publications, data are visually depicted in figures or tables. The original data behind the figures – the 'source data' – however are almost never available in a structured format that would make them findable and reusable. To address this issue, SourceData (http://sourcedata.embo.org) has built a suite of tools to capture the structure of published research data and to make published research papers discoverable based solely on their data content

The SourceData annotation system is designed to add semantic enrichment to figures and figure legends of published manuscripts. The annotation procedure consists of tagging terms that represent biological entities in figures and figure legends and assigning them a *type* (gene, protein, small molecule, cellular component, cell type, tissue, species) and a *role* in the described experimental assay. The two major roles are "assayed" and "intervention/perturbation". In a usual experimental design, the system is perturbed with an experimental controlled intervention on one or several of its components (dosage variation, gene knockout,...) and the response to this perturbation is measured at the level of assayed components. Other components of the system could have a role of normalisation (loading control) or reporter (GFP). This type of description enables to distinguish experiments that test the effect of X on Y (for example the effect of insulin on glucose) from experiments testing the effect of Y on X (eg glucose on insulin). Moreover, it allows to link two experiments together; if a first experiment tests the effect of a perturbation X on an entity Y and a second experiment tests the effect of perturbing Y on the response Z, it suggests that the data resulting from these 2 experiments might be combined to assess the potential effect of X on Z. This type of inference is facilitated by assigning standard identifiers (normalisation) from reference resources and ontologies (UniprotKB, ChEBI, Uberon, NCBI taxonomy...) to tagged terms.

Several automatic text mining tools address the process of tag identification and tag normalisation. SourceData helps the curator in this process, but this activity is still time consuming. In this work, we would like to integrate the results of PubTator (https://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/PubTator/) annotations in our pre-tagging procedure.

A preliminary study conducted on ~8000 annotated panels clearly highlighted the potential added value of pre-tagging figure legends with PubTator. However, it also showed that some improvements could be performed, and for this, we would like to take the opportunity of working together with members of the PubTator group (Zhiyong Lu and colleagues).

SourceData has now manually annotated more than 15'000 experiments (panels). This represents a unique training set for benchmarking and fine-tuning entity recognition softwares, like PubTator. The parameters of PubTator could be optimized for working at the level of figure legends. We also would like to define and implement mechanisms to automatically import results from PubTator analysis in the SourceData curation process.

We expect that this preprocessing of figure legends will dramatically shorten the time required to curate a publication in SourceData, which could then motivate potential users to adopt this system to structure and annotate published experimental datasets which will lead to an increase of linked data in the biological literature.