# CCO, a paradigm for knowledge integration

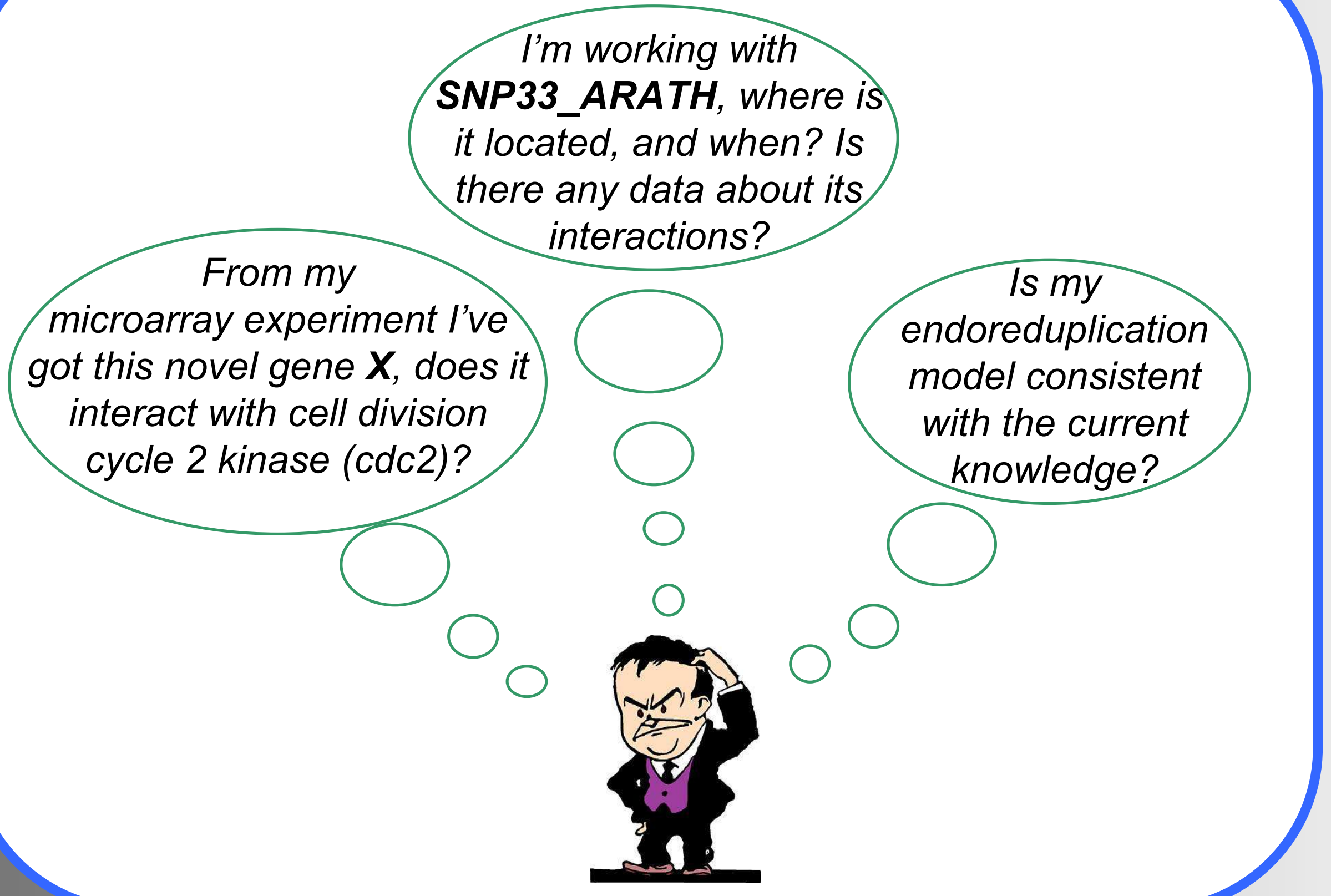
Erick Antezana[1], Vladimir Mironov[1], Mikel Egaña[2], and Martin Kuiper[1]

[1] VIB Dept. of Plant Systems Biology, Ghent University. Technologiepark 927, B-9052 Ghent BELGIUM. E-mail: {erant,vlmir, makui}@psb.ugent.be

[2] The University of Manchester. School of Computer Science, Oxford Road, M13 9PL, UK. E-mail: mikel.eganaaranguren@cs.man.ac.uk

http://www.CellCycleOntology.org

## Motivating scenarios



I'm working with **SNP33_ARATH**, where is it located, and when? Is there any data about its interactions?
From my microarray experiment I've got this novel gene **X**, does it interact with cell division cycle 2 kinase (cdc2)?
Is my endoreduplication model consistent with the current knowledge?

## Objective

To capture the knowledge about the **cell cycle** process (particularly its dynamic facets) and to promote sharing, reuse and enable better computational integration with existing resources (semantic web). The ultimate aim is to support evaluation and generation of hypotheses via reasoning services about cell-cycle regulation.

**Target organisms:** *S. cerevisiae, S. pombe, A. thaliana and H. sapiens*.

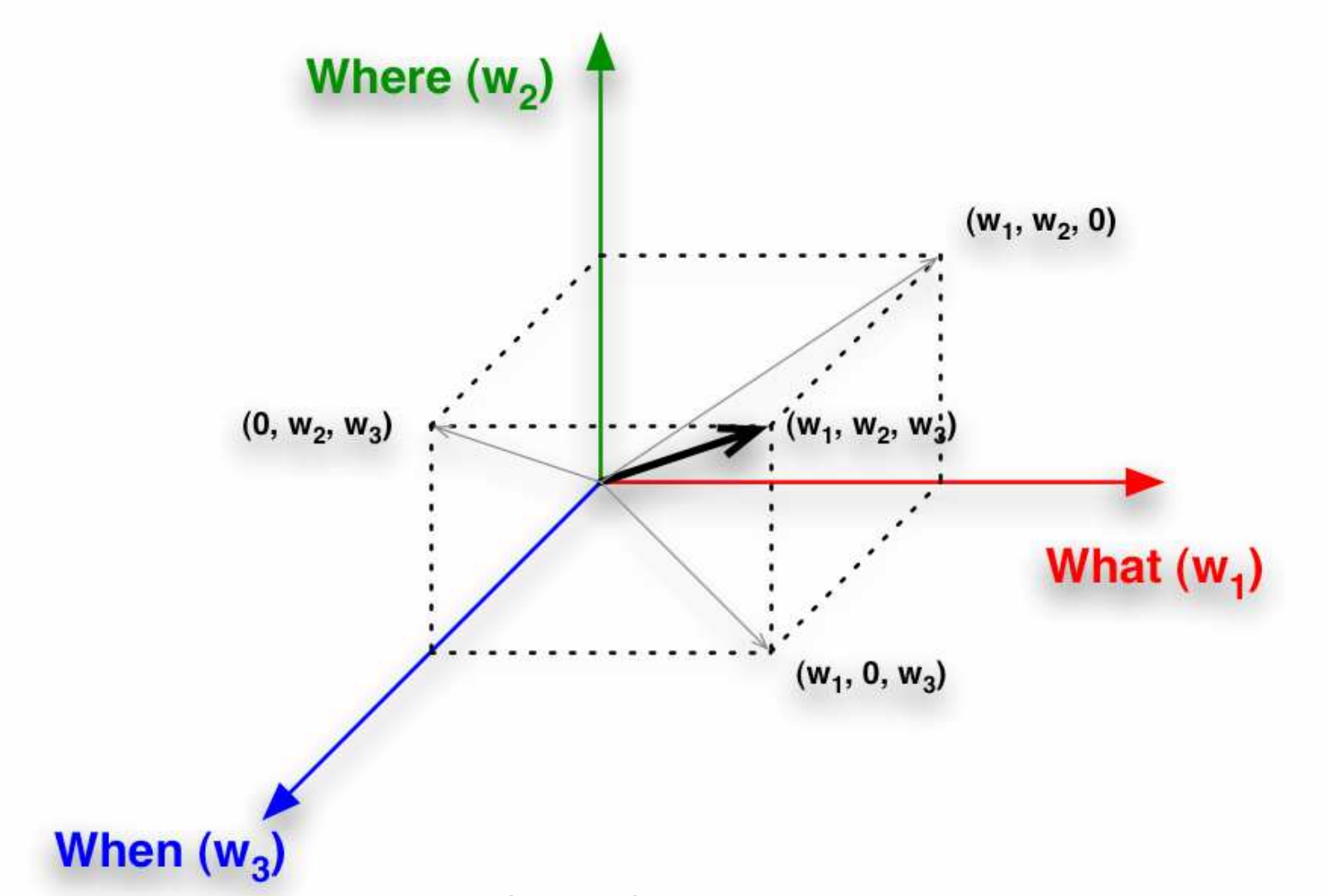


*Fig 1. W3 paradigm.*

W3 paradigm: **What-Where-When**. Sample piece of knowledge: *« Cyclin B (w1) is located in Cytoplasm (w2) during Interphase (w3) ».*

## Pipeline

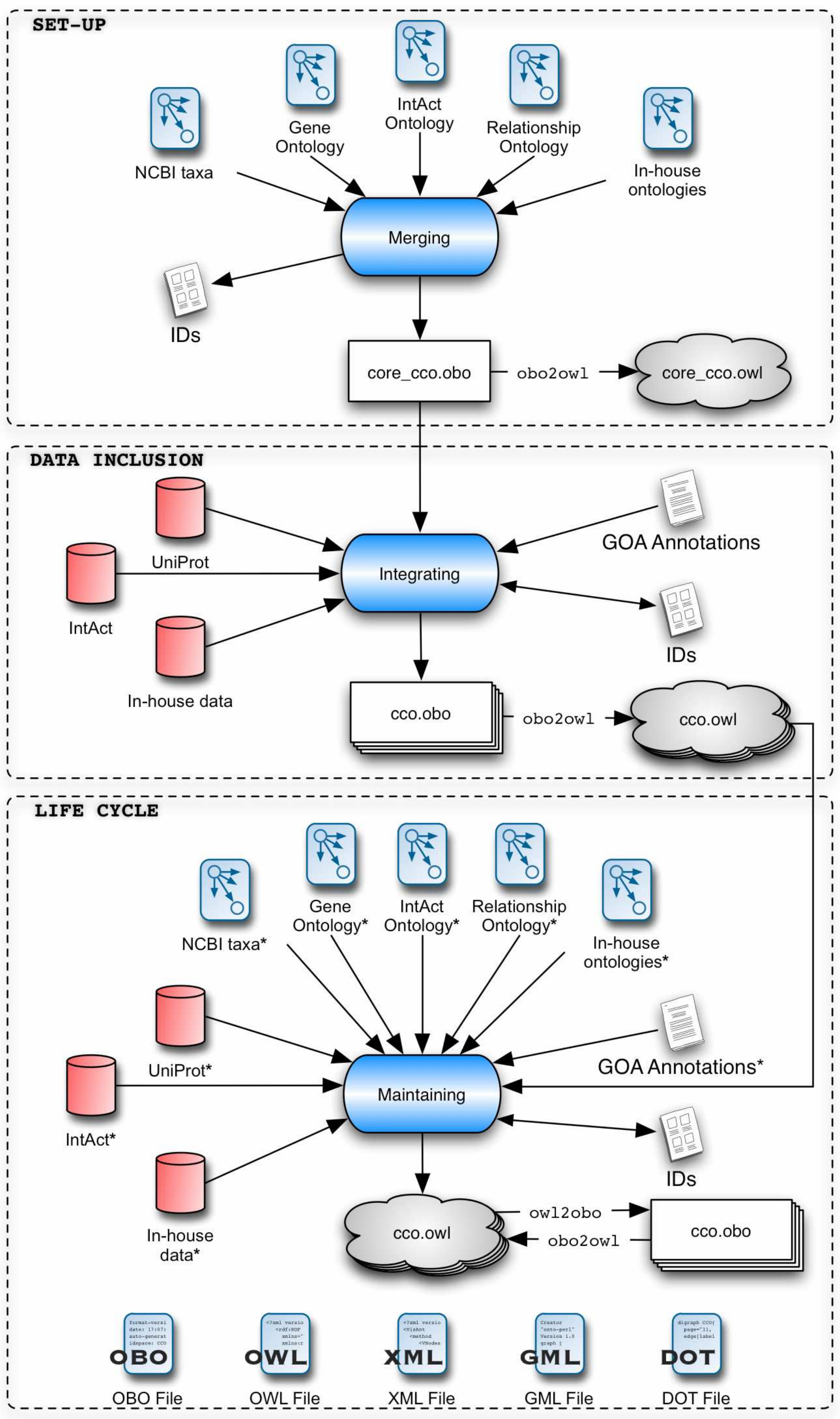


*Fig 2. Data integration pipeline.*

## Exploring CCO

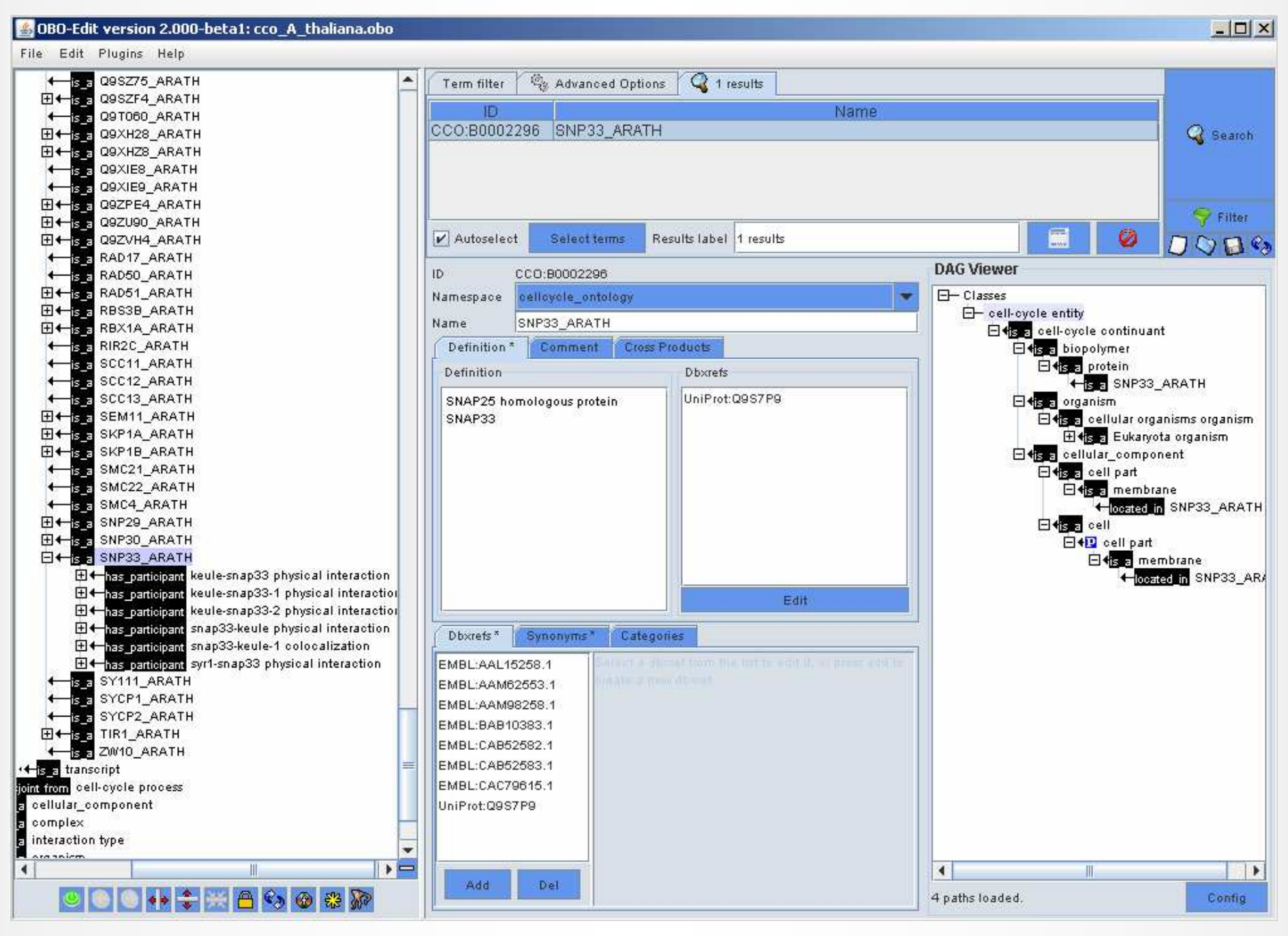


*Fig 3. CCO in OBO-Edit.*

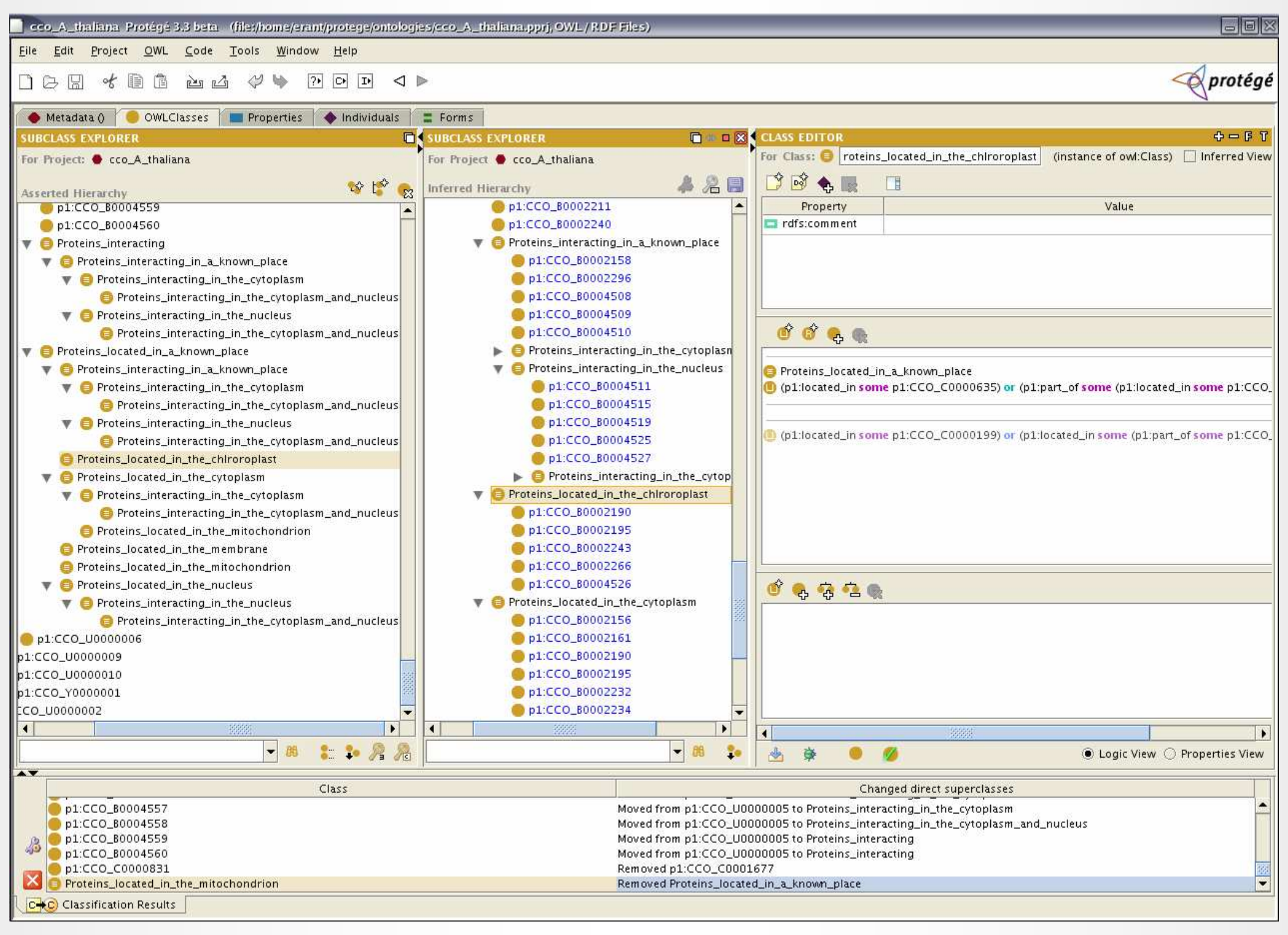
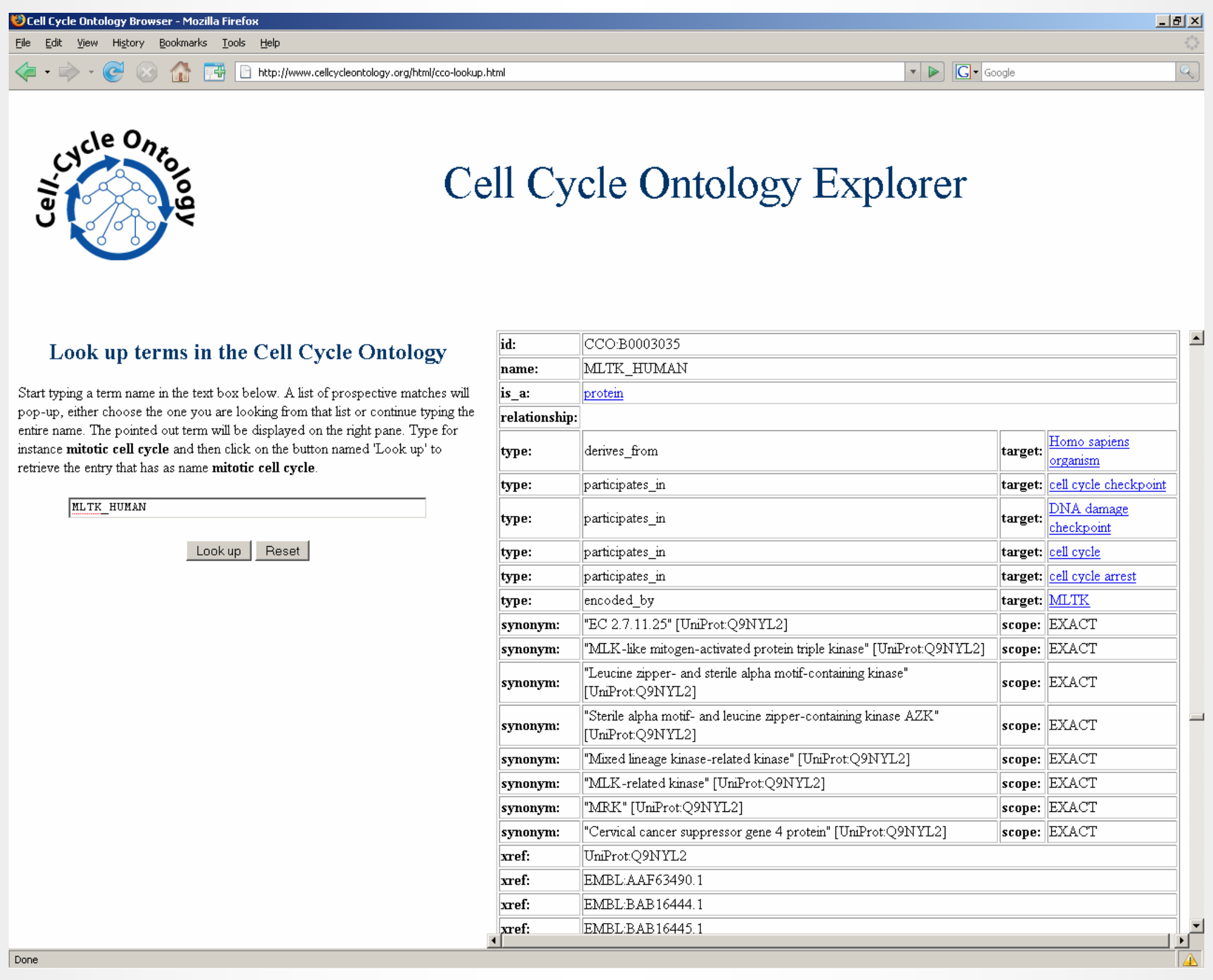


*Fig 4. CCO in Protégé.*



Cell Cycle Ontology Browser - Mozilla Firefox

http://www.cellcycleontology.org/html/cco-lookup.html

Cell Cycle Ontology Explorer

Look up terms in the Cell Cycle Ontology

Start typing a term name in the text box below. A list of prospective matches will pop-up, either choose the one you are looking from that list or continue typing the entire name. The pointed out term will be displayed on the right pane. Type for instance **mitotic cell cycle** and then click on the button named 'Look up' to retrieve the entry that has as name **mitotic cell cycle**.

MLTK_HUMAN

Look up · Reset

| | | | |
|---|---|---|---|
| **id:** | CCO:B0003035 | | |
| **name:** | MLTK_HUMAN | | |
| **is_a:** | protein | | |
| **relationship:** | | | |
| **type:** | derives_from | **target:** | Homo sapiens organism |
| **type:** | participates_in | **target:** | cell cycle checkpoint |
| **type:** | participates_in | **target:** | DNA damage checkpoint |
| **type:** | participates_in | **target:** | cell cycle |
| **type:** | participates_in | **target:** | cell cycle arrest |
| **type:** | encoded_by | **target:** | MLTK |
| **synonym:** | "EC 2.7.11.25" [UniProt:Q9NYL2] | **scope:** | EXACT |
| **synonym:** | "MLK-like mitogen-activated protein triple kinase" [UniProt:Q9NYL2] | **scope:** | EXACT |
| **synonym:** | "Leucine zipper- and sterile alpha motif-containing kinase" [UniProt:Q9NYL2] | **scope:** | EXACT |
| **synonym:** | "Sterile alpha motif- and leucine zipper-containing kinase AZK" [UniProt:Q9NYL2] | **scope:** | EXACT |
| **synonym:** | "Mixed lineage kinase-related kinase" [UniProt:Q9NYL2] | **scope:** | EXACT |
| **synonym:** | "MLK-related kinase" [UniProt:Q9NYL2] | **scope:** | EXACT |
| **synonym:** | "MRK" [UniProt:Q9NYL2] | **scope:** | EXACT |
| **synonym:** | "Cervical cancer suppressor gene 4 protein" [UniProt:Q9NYL2] | **scope:** | EXACT |
| **xref:** | UniProt:Q9NYL2 | | |
| **xref:** | EMBL:AAF63490.1 | | |
| **xref:** | EMBL:BAB16444.1 | | |
| **xref:** | EMBL:BAB16445.1 | | |

Done

*Fig 5. CCO Explorer (online).*

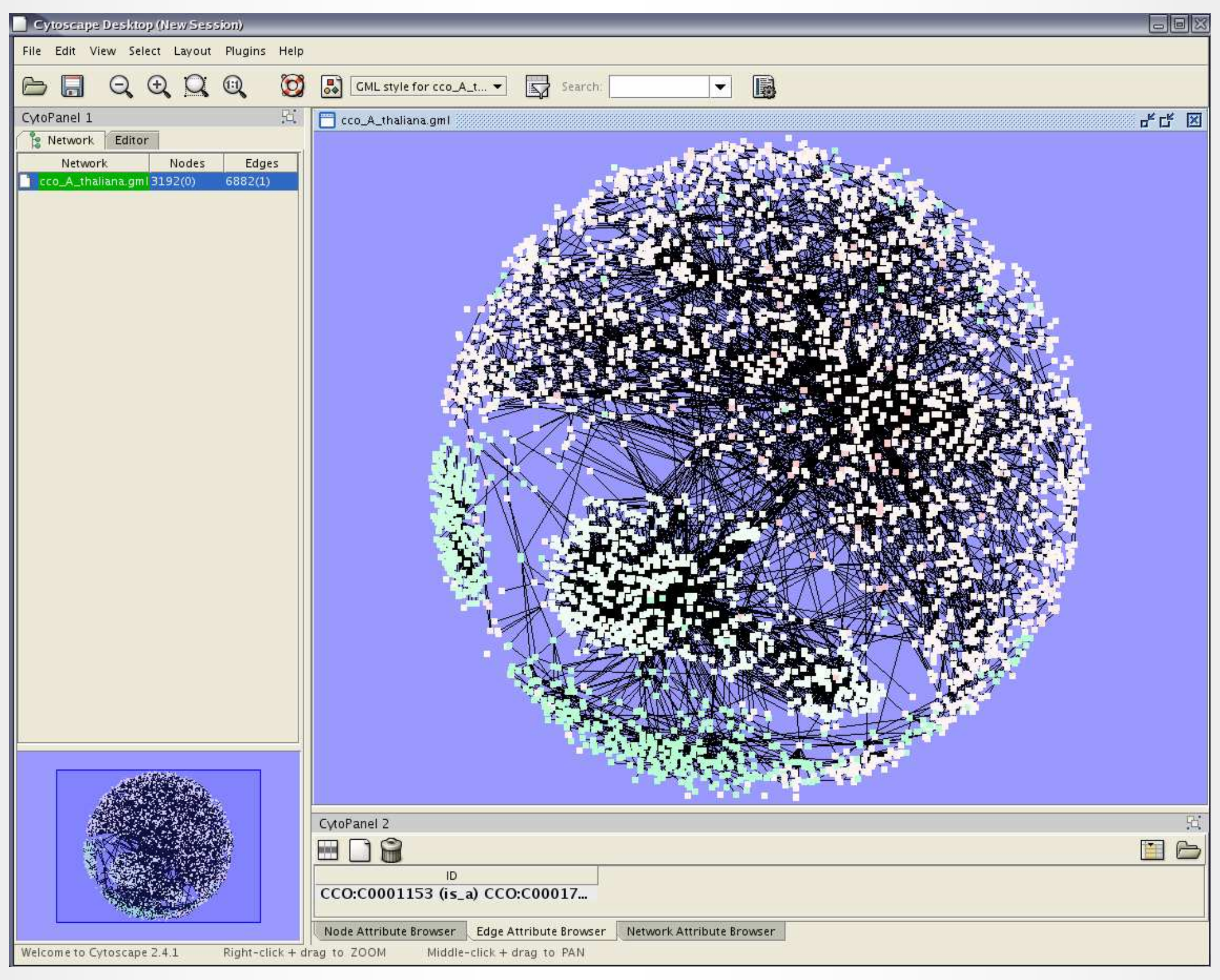


*Fig 6. CCO in Cytoscape.*

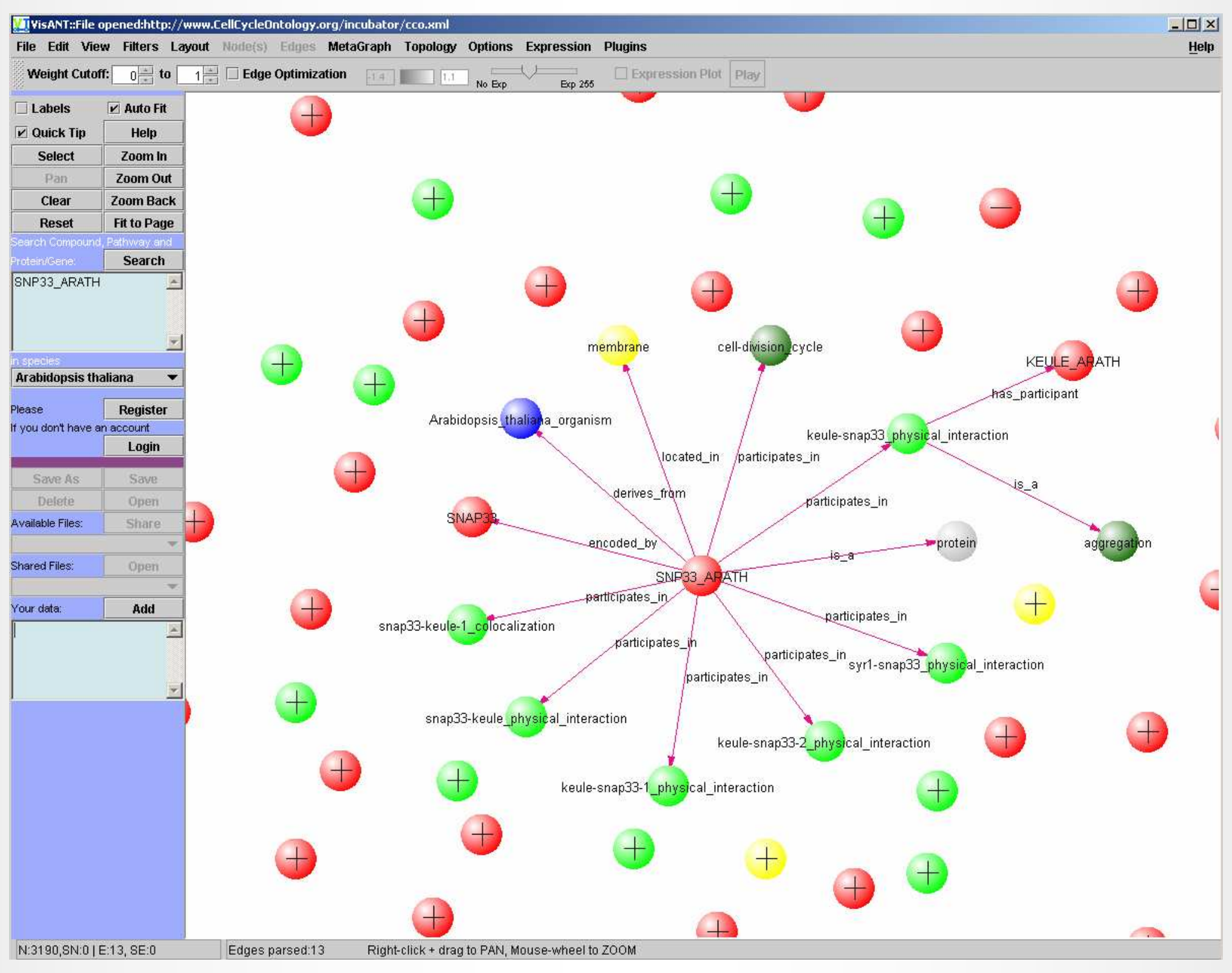


*Fig 7. CCO in visANT.*

## Method

CCO should capture the semantics and spatio-temporal relationships (*Fig 1*) of cell-cycle components (proteins, genes, cellular locations, phases, and so on).

Data sources:

- GO (CL, CC branches)
- RO
- MI (IntAct ontology)
- GOA files
- PPI: IntAct
- NCBI taxonomy
- Cell cycle functional data
- Data obtained with bio-tools (e.g.OrthoMCL)

OBO and OWL-DL formats have been chosen for representing the knowledge. RACER is mainly used for checking the data consistency and for doing classifications.

## Reasoning results

- There are a number of relationships in GO (core source of CCO) that might have been better annotated as *part_of* instead of *is_a*.
- The results inspired the GO team to made some amendments to the process part of the GO (e.g regulation of cell cycle).
- Inconsistencies found in the data about the cellular localizations and protein-protein interactions.

## Conclusions and Results

- Fully automated data integration pipeline (nightly launched) was developed (*Fig 2*).
- Concrete problems and results related to the implementation of automatic format mappings (OWL, XML, DOT, GML) between ontologies and inconsistency checking issues have been identified.
- Exports in several formats developed (*Fig 3-7*).
- Existing integration obstacles due to the diversity of data formats and lack of formalization approaches as well as the trade-offs that are common in biological sciences.

## Future work

- Knowledge will be weighted (e.g. evidence codes) expressing the support media similar to those implemented in GO (experimental, electronically inferred, and so forth).
- Ontolome analysis (e.g. hypothesis generation by ontology alignments).
- Advanced query system will be developed (DL-based).
- More data to be integrated (upon feedback).

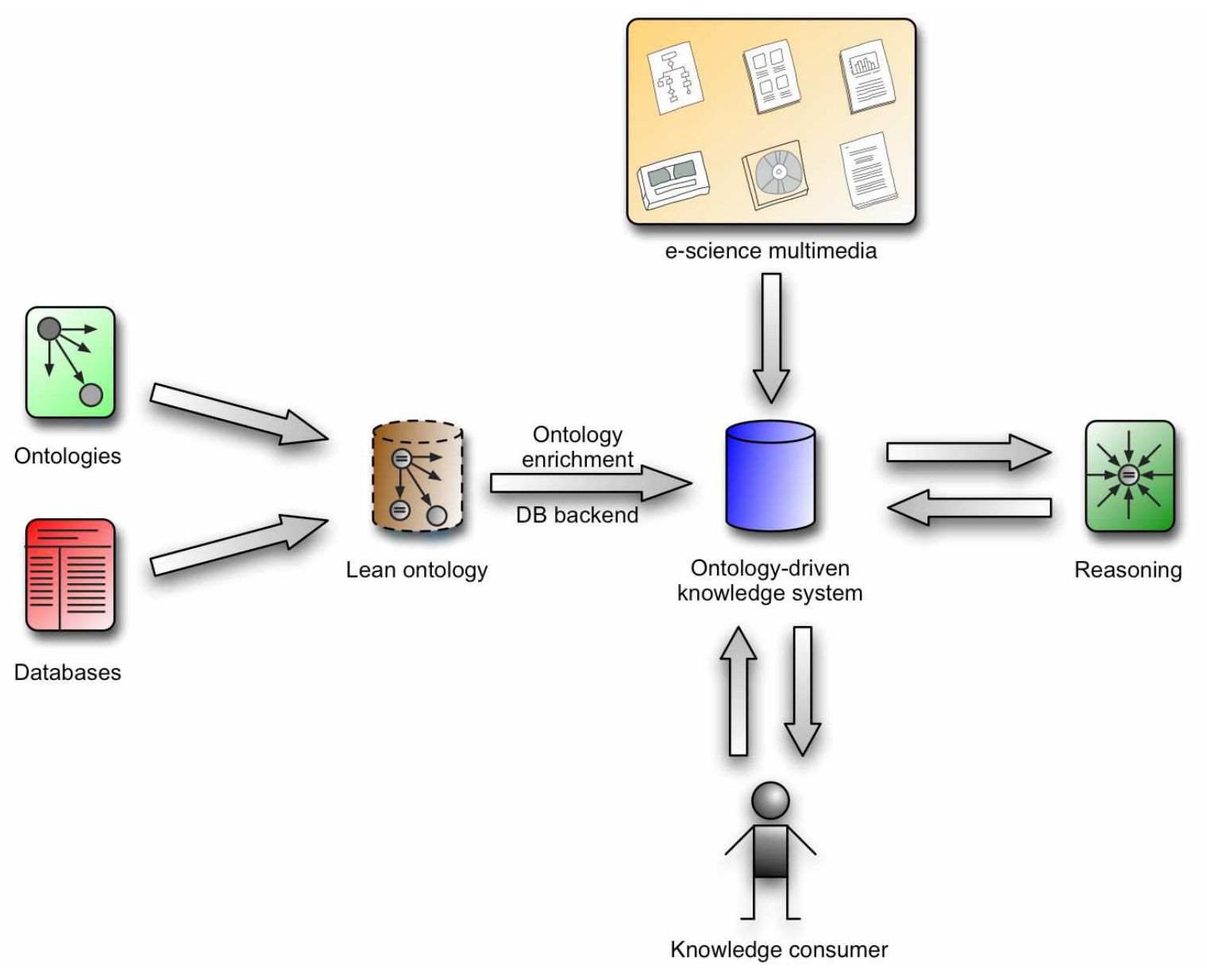


*Fig 8. Outlook into the future*

## Acknowledgements

This work has been funded by the EU (contract number LSHG-CT-2004-512143), The Manchester University, EPSRC, and Marie Curie EST.