FLOW CYTOMETRY & ONTOLOGIES

Mélanie Courtot

BC Cancer Agency, Vancouver, Canada

Immunology Ontologies and Their Applications in Processing Clinical Data, Buffalo, June 12th 2012

FLOW CYTOMETRY IN THE ONTOLOGY FOR BIOMEDICAL INVESTIGATIONS (OBI)

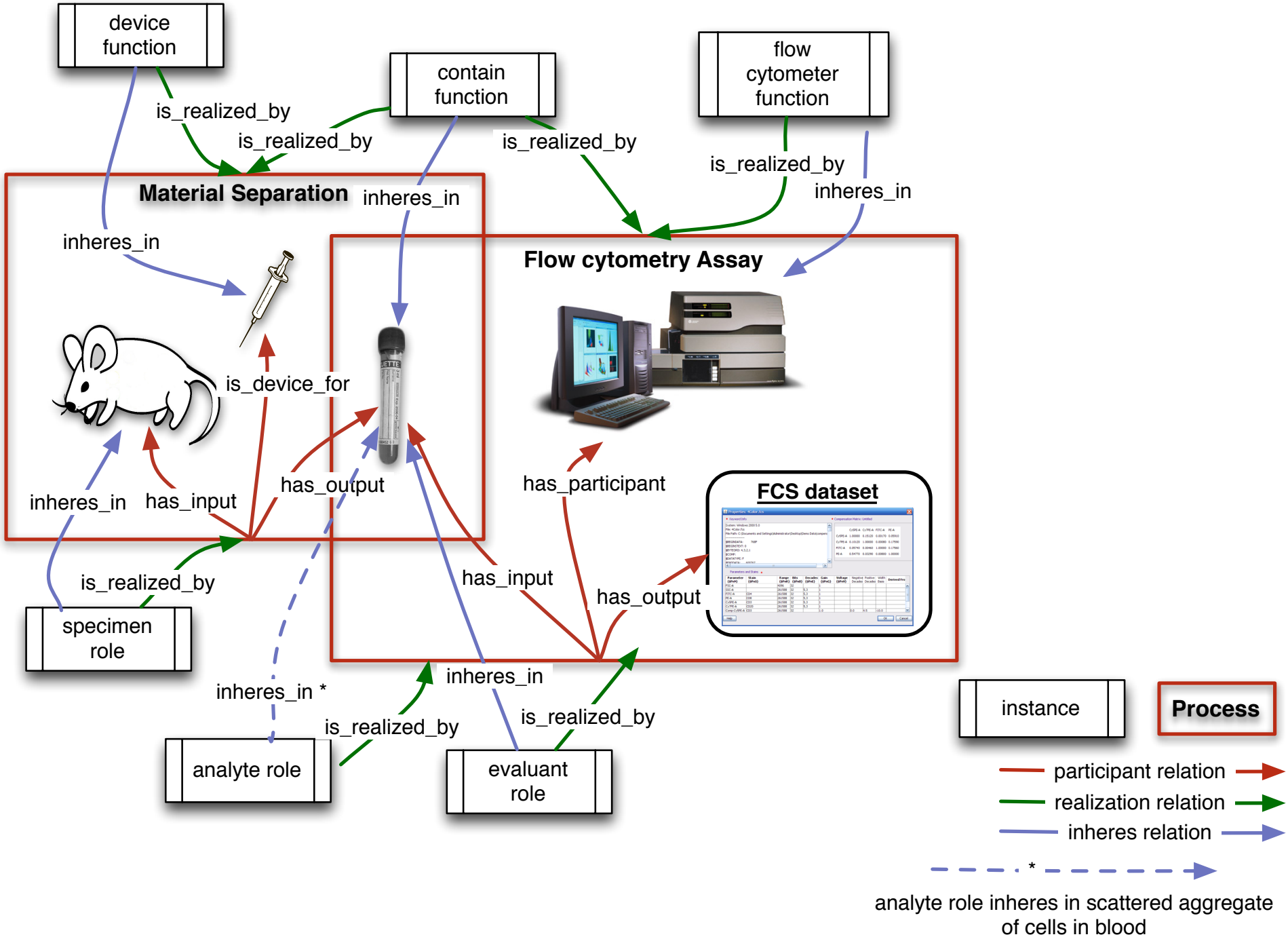
Mélanie Courtot[1], Ryan Brinkman[1] and the OBI Consortium[2]

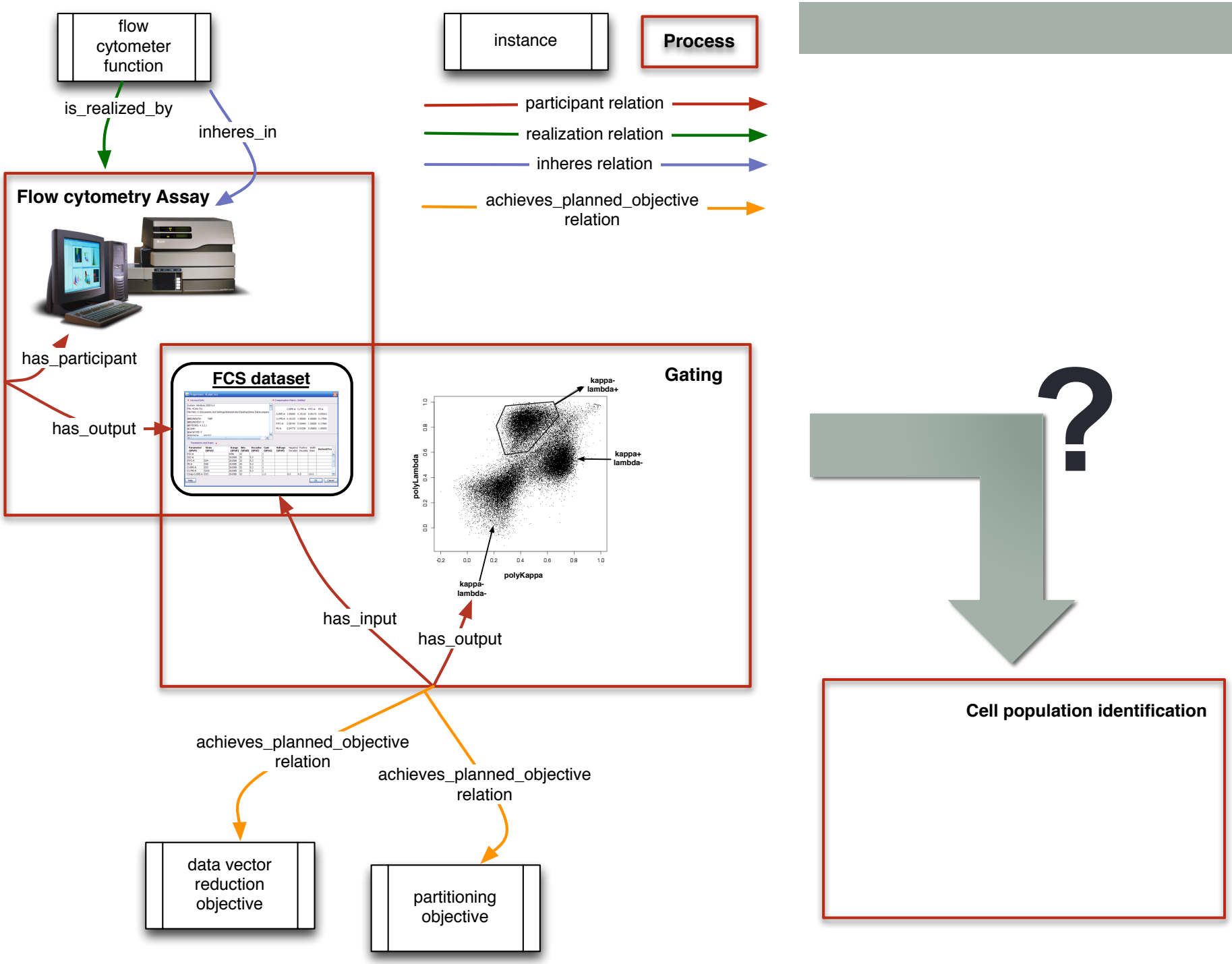
[1] BC Cancer Agency, Vancouver, Canada

[2] http://purl.obolibrary.org/obo/obi

OBI terms

- Instruments and parts
  - Flow cytometers sorters, analyzers, light sources, filters….
- Fluorochromes (in the Chemical Entities of Biological Interest (ChEBI))
  - More than 400 have been added, each including formula, synonyms…
- Processes
  - Flow cytometry assays
  - Gating
- Processes objectives
  - Partitioning


device function
is_realized_by
contain function
is_realized_by
is_realized_by
flow cytometer function
is_realized_by
inheres_in
Material Separation
inheres_in
inheres_in
Flow cytometry Assay
is_device_for
inheres_in
has_input
has_output
has_participant
FCS dataset
is_realized_by
specimen role
has_input
has_output
inheres_in
inheres_in *
is_realized_by
is_realized_by
analyte role
evaluant role
instance
Process
participant relation
realization relation
inheres relation
*
analyte role inheres in scattered aggregate of cells in blood


flow cytometer function
is_realized_by
inheres_in
instance
Process
participant relation
realization relation
inheres relation
achieves_planned_objective relation
Flow cytometry Assay
has_participant
has_output
FCS dataset
Gating
kappa- lambda+
kappa+ lambda-
kappa- lambda-
polyLambda
polyKappa
has_input
has_output
achieves_planned_objective relation
achieves_planned_objective relation
data vector reduction objective
partitioning objective
?
Cell population identification

Acknowledgements

- Elizabeth Goralczyk, John Quinn and Josef Spidlen
- Ryan Brinkman, Richard Scheuermann
- James Malone and Elisabetta Manducchi, Bjoern Peters and the OBI consortium
- The ChEBI team
- National Institutes of Health (NIH)/National Institute of Biomedical Imaging and Bioengineering (NIBIB) funding

CONNECTING FCM ANALYSIS RESULTS WITH THE CELL ONTOLOGY IN AN AUTOMATED WAY

Adrin Jalali[1], **Mélanie Courtot**[1], Raphael Gottardo[2], Richard Scheuermann[3], Ryan Brinkman[1]

[1]BC Cancer Agency, Vancouver, Canada
[2]Fred Hutchinson Cancer Research Center, Seattle, US
[3]J. Craig Venter Institute, San Diego, US

Automated methods can't semantically label cell populations

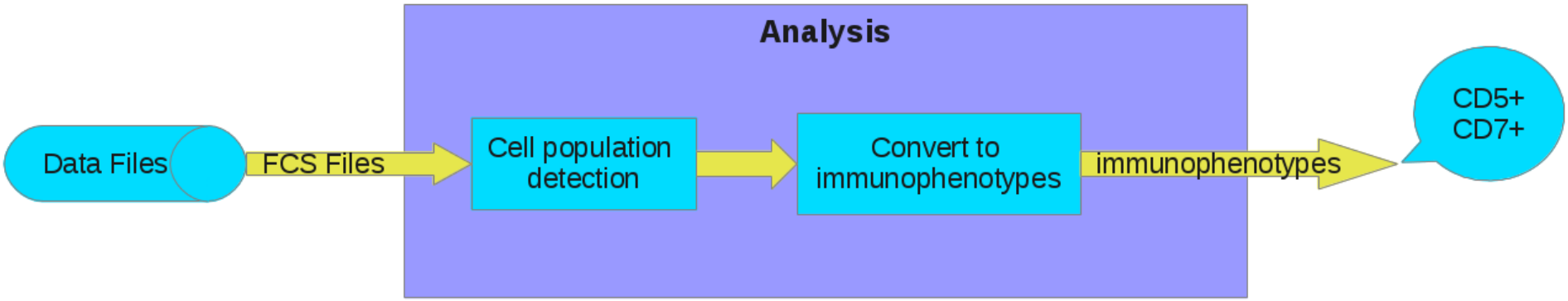
- Different researchers refer to cell populations types using different labels, depending on the experiment context
- Automated analysis label cell populations via their belonging to a cluster
- Cell groups can be identified via different immunophenotype, e.g., CD5+ or CD7+ to identify T-cells
- As a result, outputs from different sources can not be compared
- Labeling cell populations using common natural language will facilitate comparison and collaboration

SOLUTION

- A framework allowing to label immunophenotypes resulting from a Flow Cytometry (FCM) analysis (automated or manual) to a consensus label will allow researchers to unambiguously refer to a defined cell population
- Previous research on the same cell population and/or related will become accessible, even if different markers were used
- We call this **the Cell Population Labeler (CPL)**

FCM analysis

- Analysis outputs an immunophenotype (i.e., a set of markers, such as CD3+CD4+)
- Markers can be present/absent, or present at various levels such as low, intermediate and high
- Output fed to the CPL



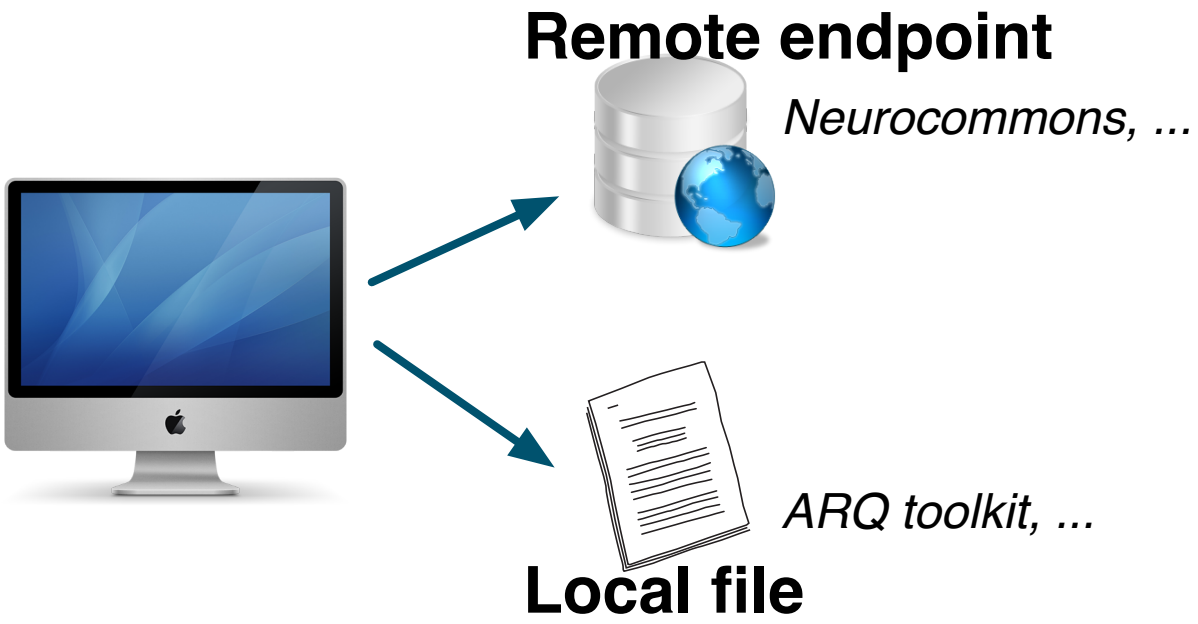
Cell Population Labeler (CPL)

- Identifies a subset of the Cell Ontology (CL) tree as corresponding to the given immunophenotype
- This subset can be a single node, empty, or a sub-tree
- With increase in the immunophenotype specificity, we expect the sub-tree (DAG) to get progressively pruned, and ideally to retrieve only a single node

Step 1: Access to the CL

- Use SPARQL to query the CL OWL

  Select ?celllabel where { ?x a owl:Class.?x rdfs:label ?celllabel.}



Problem: Available CL needs to be updated

Step 2: browse the content of the CL

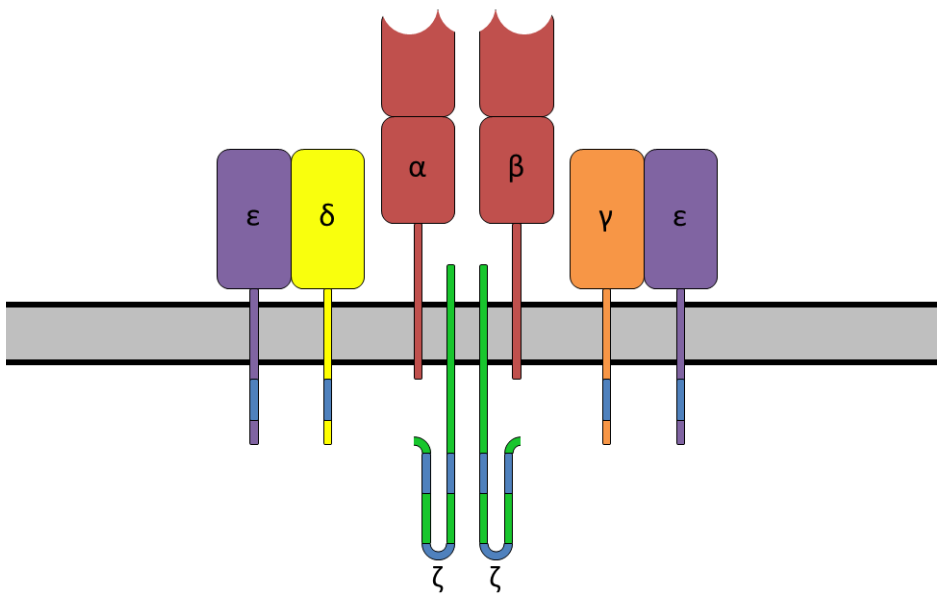
- Minor issues
  - Modelization issues, e.g., properties hierarchy
  - Missing terms
  - Release artifacts, e.g. duplicated relations

- Feedback from users to improve the current resource.
- CL tracker available and CL team very responsive.

http://sourceforge.net/tracker/?group_id=76834&atid=925065

Step 2: Browse the content of the CL

- Missing information, such as parthood relationship between receptor and subunits. Need coordination between different resources (e.g., Gene Ontology, Protein Ontology)



The T-cell receptor complex with TCR-α and TCR-β chains (top), ζ-chain accessory molecules (bottom) and CD3 (represented byCD3γ, CD3δ and two CD3ε).

Source: wikipedia

Step 2: browse the content of the CL

- Scope of the CL

- CL aims at identify those markers that are necessary and sufficient to define a cell type.
- Some work in Richard's group to list extra marker expression characteristics for hematopoeitic cell types.

Step 3: Implementation of an automated pipeline

- Pipeline in R
- R is a free, open source, robust statistical programming environment for Windows, Mac & Linux that offers a wide range of statistical and visualization methods
- BioConductor provides R software modules for biological and clinical data analysis
- Integrates with other software tools via open data standards
- Use SPARQL from R
  - Several libraries available, need investigation/testing

28+ R packages for Flow Analysis (all since 2007)

- **Data processing & Visualization**
  - flowCore: Read/write & process flow data
  - plateCore: Analyze multiwell plates
  - flowUtils: Import gates, transformation and compensation
  - flowStats: Advanced statistical methods and functions
  - ncdfFlow: Advanced methods for large dataset processing
  - flowQ: Quality control of ungated data
  - QUALIFIER: Quality control and assessment of gated data
  - flowViz: Visualization (e.g., histograms, dot plots, density plots)
  - flowPlots: Graphical displays with statistical tests
  - flowWorkspace: Importing FlowJo workspaces
  - iFlow: GUI for exploratory analysis and visualization
  - flowTrans: Estimates parameters for data transformation
- **Gating**
  - flowClust: Clustering using t-mixture model with Box-Cox transformation
  - flowMerge: flowClust + entropy-based merging
  - flowMeans: k-means clustering and merging using the Mahalanobis distance
  - SamSpectral: Efficient spectral clustering using density-based down-sampling
  - flowPeaks: Unsupervised clustering using k-means \& mixture model
  - flowFP: Fingerprint generation
  - flowPhyto: Analysis of marine biology data
  - flowQB: Q&B analysis
  - FLAME: Multivariate finite mixtures of skew and heavy-tailed distributions
  - flowKoh: Self-organizing maps
  - NMF-curvHDR: Density-based clustering and non-negative matrix factorization
  - flowCore-flowStats: Sequential gating and normalization and a Beta-Binomial model
  - PRAMS: 2D Clustering and logistic regression
  - SPADE: Density-based sampling, k-means clustering, and minimum spanning trees
- **Discovery**
  - flowType: Automated phenotyping using 1D gates extrapolated to multiple dimensions
  - RchyOptimyx: Cellular hierarchies correlated with outcome of interest

Known limitations

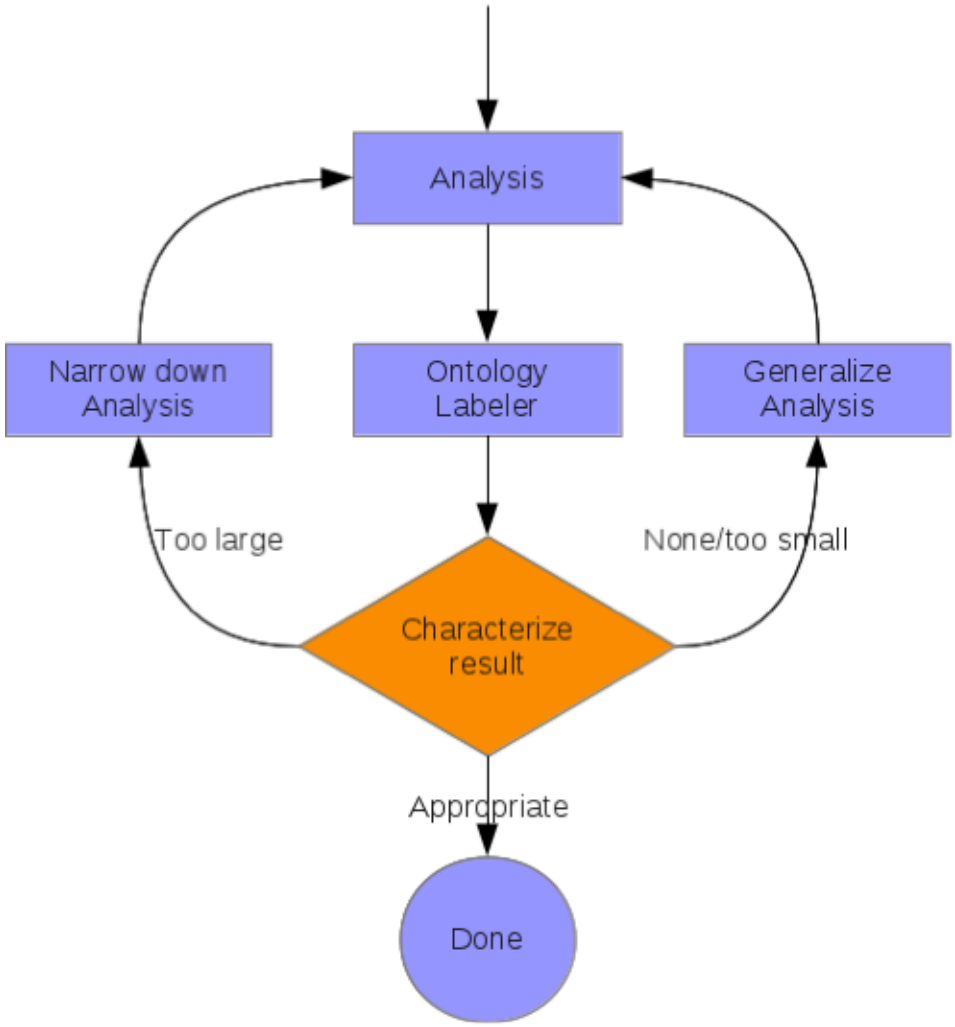
- Uses string matching between output immunophenotypes and CL markers
- It will be challenging to account for all lexical variants
  - CD45RO+ <-> has_plasma_membrane_part some 'receptor-type tyrosine-protein phosphatase C isoform CD45RO'
  - Relations: lacks_plasma_membrane_part, has_low_plasma_membrane_amount…
  - Synonyms: T cell, T-cell…

Proposal - flowCL

- A small extension to the CL
- Build specifically to address our use case
- Would allow for flexibility in development
- As it would import CL, it could easily be incorporated if desired, or distributed as distinct extension

Result

- Returned result can be refined with addition of additional markers
- Ideal case: single node



Additional features

- If multiple phenotypes, increase the degree of confidence of the result with the number of returned result sets it belongs too
- Based on the analysis output (e.g., if we have a DAG) we can exploit this hierarchy to favor results matching more specific immunophenotypes.
  - Prefer the value T helper lymphocytes over T cell lymphocytes when the immunophenotype is CD3+CD4+

Summary – current issues

- **String matching** between immunophenotypes and cell populations
- How to deal with **relative abundance** of markers (dim/bright)
- On the analysis side, how to identify population based on **previous knowledge** (e.g., kappa-lambda+)
- **Access** to the CL: remote, local, both?
- **Tooling** evaluation
- **CL content**: ensure action items are ported to release. Coordination with other efforts. Scope: cell knowledgebase?

Acknowledgements

- Adrin Jalali, Raphael Gottardo, Richard Scheuermann, Ryan Brinkman
- Alexandre Diehl and the CL developers
- National Institutes of Health (NIH)/National Institute of Biomedical Imaging and Bioengineering (NIBIB) funding