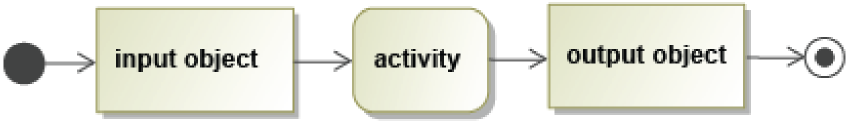
Biocuration workflow catalogue

This document is a first attempt to develop an intuitive knowledge engineering methodology biocurators to describe their workflows in a generic, human readable format. This format may then be transcribed into a computational form. The basic idea is to encourage biocurators to describe a semi-formalized narrative where the inputs, process and outputs of each step of the workflow are explicitly described. This approach is intuitively based on the way that scientific protocols are described.

We use the outlining methodology of a word-processing document to provide the necessary indexing between parts of the workflow.

The crucial element to understand about this formulation that the workflow is (conceptually) a **bipartite graph** (a graph consisting of two types of nodes where the nodes of one type only connect to nodes of the other type), made up of **activities** and **data-objects**. These are described graphically in the following way:



Within the graphical workflows, if we were to include every input and output object of every activity, the graphs would become cluttered, overly busy and difficult to read. Therefore, the first time a data object is either created or used, it is represented explicitly. Unless a different object is created within the workflow (not if the current object is only modified, annotated or edited), we only represent the first object explicitly.

We use the framework provided by Activity Diagrams in the Universal Modeling Language to structure these workflows (see http://en.wikipedia.org/wiki/Activity_diagram for a preliminary introduction).

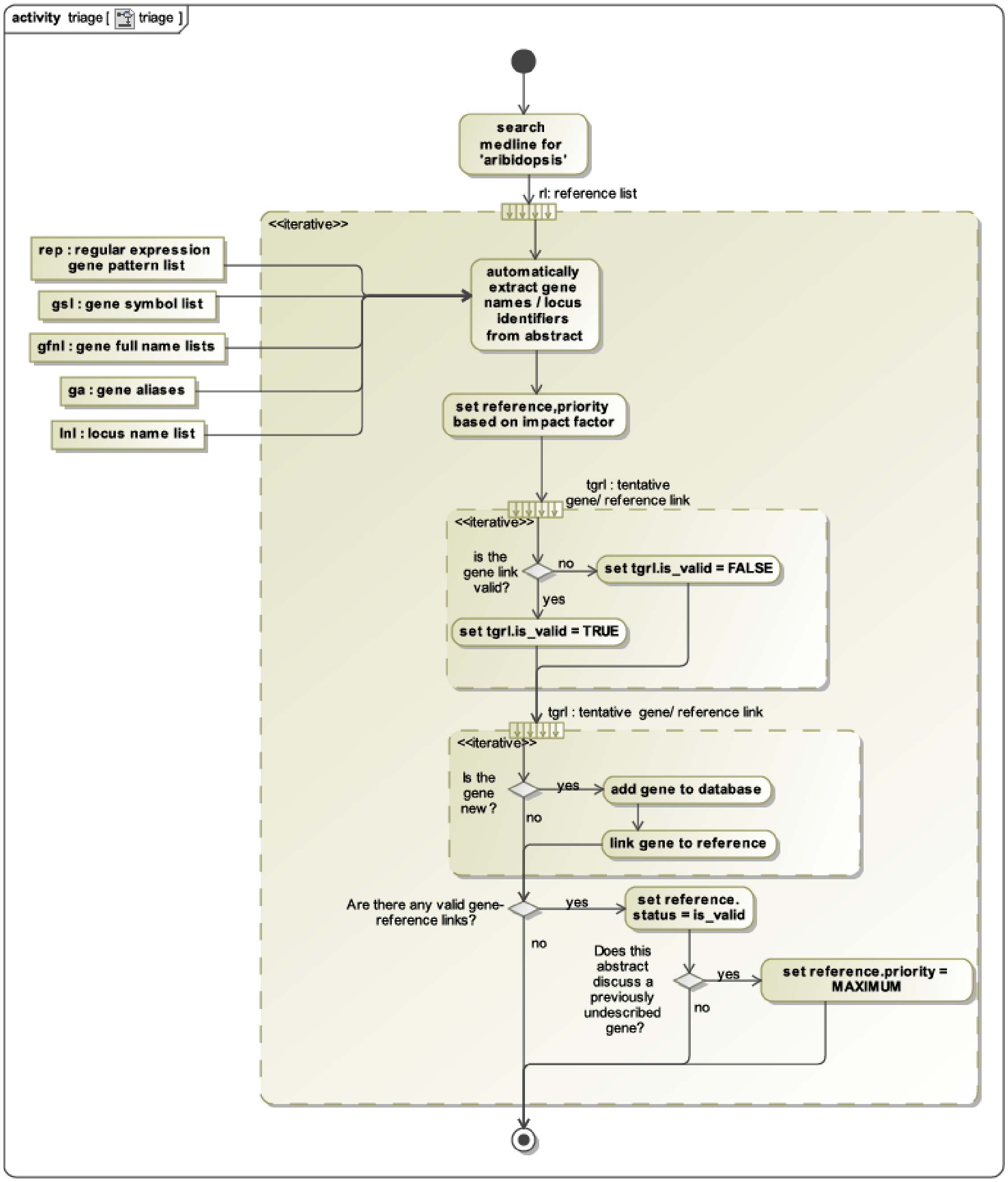
We have attempted to capture a (somewhat) representative an example set of biocuration workflows as UML Activity Diagrams. These are not formal models, but provide a standard graphical format to help us understand commonalities between the biocuration tasks facing different groups.

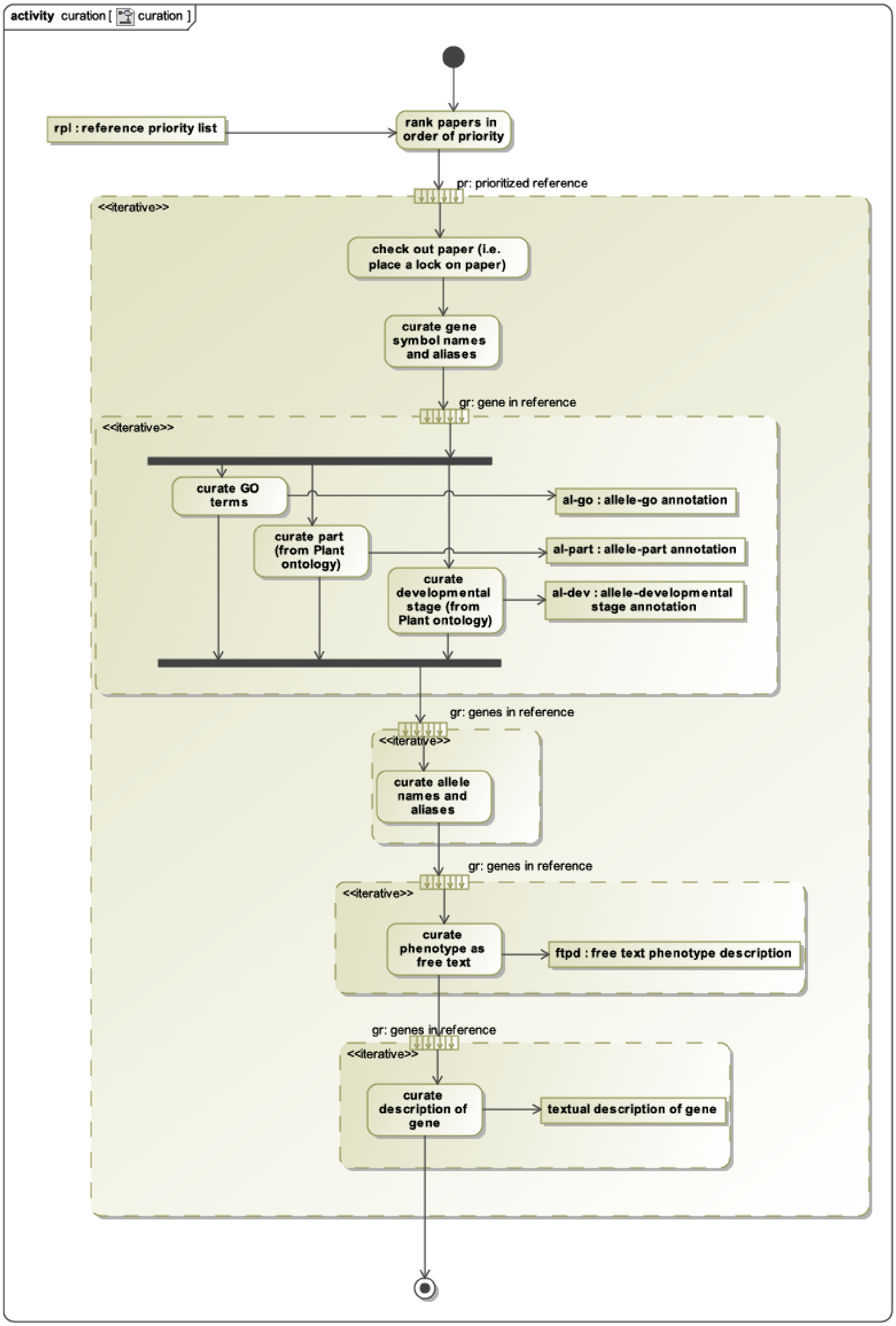
Model Organism Databases

The Arabidopsis Information Resource (TAIR)

Here we describe three workflows, all relating to different aspects of the process.

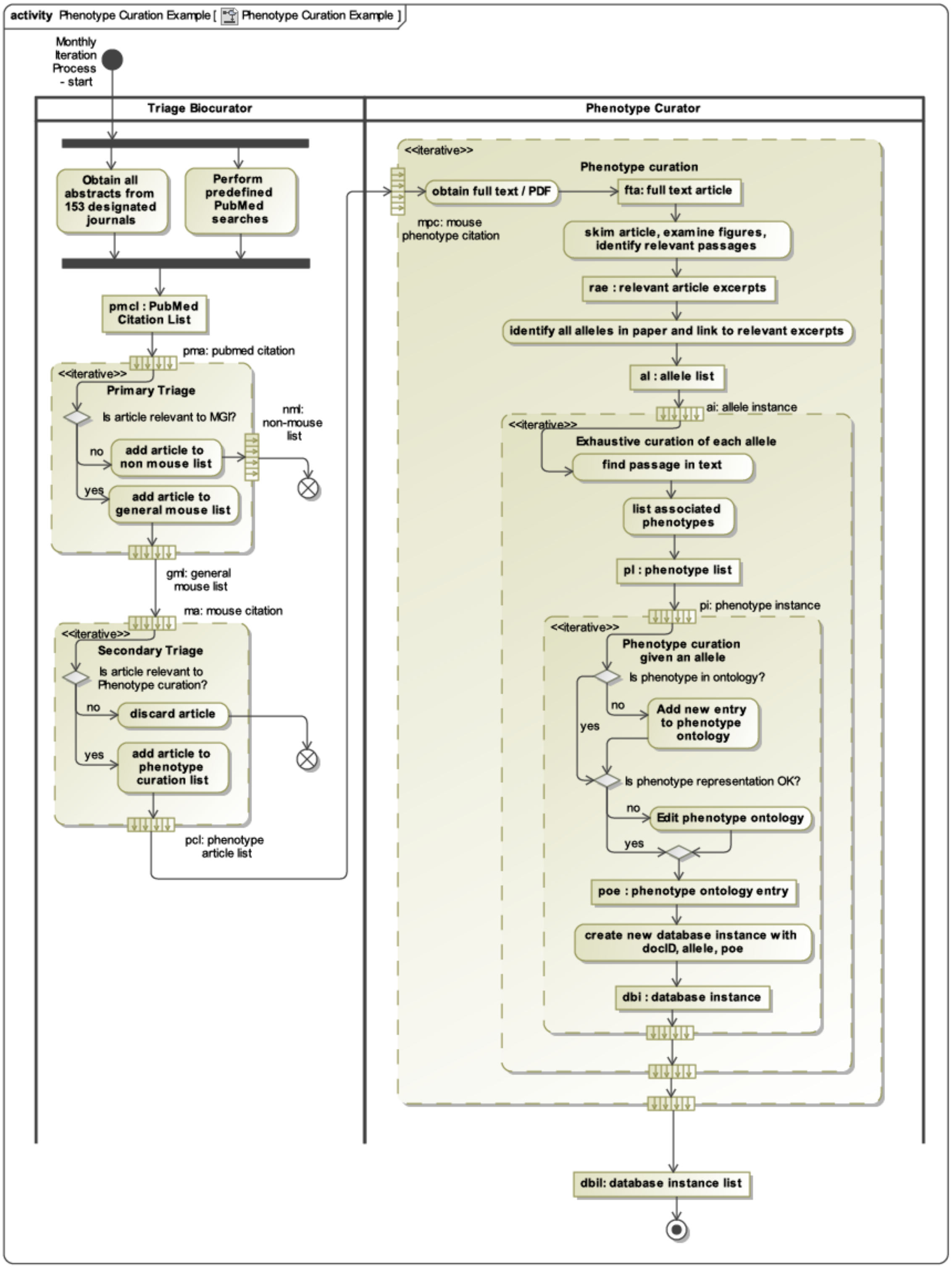
- document triage: where curators identify which papers are of interest
- curation: where curators extract relevant information from the text and input it into their local database resource.


activity triage [triage]
search medline for 'aribidopsis'
rl: reference list
<<iterative>>
rep : regular expression gene pattern list
gsl : gene symbol list
gfnl : gene full name lists
ga : gene aliases
lnl : locus name list
automatically extract gene names / locus identifiers from abstract
set reference,priority based on impact factor
tgrl : tentative gene/ reference link
<<iterative>>
is the gene link valid?
no
set tgrl.is_valid = FALSE
yes
set tgrl.is_valid = TRUE
tgrl : tentative gene/ reference link
<<iterative>>
Is the gene new ?
yes
add gene to database
no
link gene to reference
Are there any valid gene-reference links?
yes
set reference. status = is_valid
no
Does this abstract discuss a previously undescribed gene?
yes
set reference.priority = MAXIMUM
no


activity curation [curation]
rpl : reference priority list
rank papers in order of priority
pr: prioritized reference
<<iterative>>
check out paper (i.e. place a lock on paper)
curate gene symbol names and aliases
gr: gene in reference
<<iterative>>
curate GO terms
al-go : allele-go annotation
curate part (from Plant ontology)
al-part : allele-part annotation
curate developmental stage (from Plant ontology)
al-dev : allele-developmental stage annotation
gr: genes in reference
<<iterative>>
curate allele names and aliases
gr: genes in reference
<<iterative>>
curate phenotype as free text
ftpd : free text phenotype description
gr: genes in reference
<<iterative>>
curate description of gene
textual description of gene

Mouse Genome Informatics

This was a first attempt at capturing the logic of the process of annotating these workflows, and was constructed without any input from MGI curators. We include it here as a demonstration of the underlying representation and would welcome feedback from MGI staff.



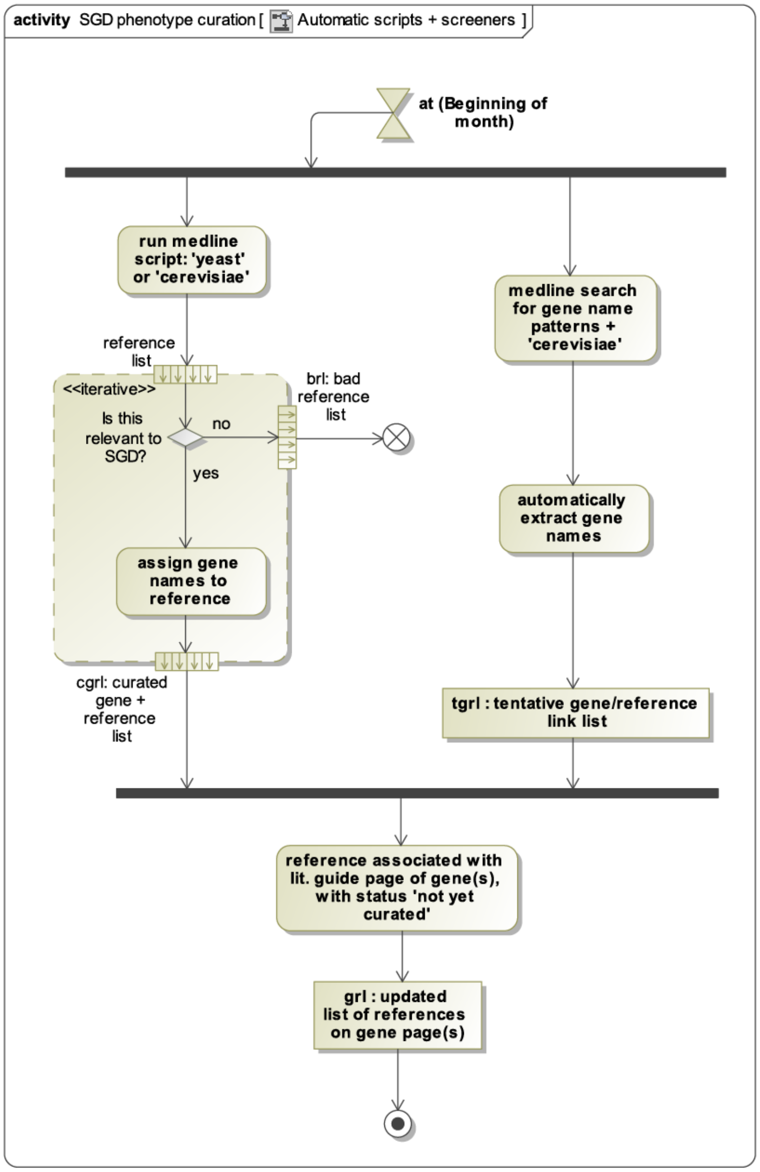
activity Phenotype Curation Example [Phenotype Curation Example]
Monthly Iteration Process - start
Triage Biocurator
Phenotype Curator
Obtain all abstracts from 153 designated journals
Perform predefined PubMed searches
pmcl : PubMed Citation List
pma: pubmed citation
<<iterative>>
Primary Triage
Is article relevant to MGI?
no
add article to non mouse list
yes
add article to general mouse list
nml: non-mouse list
gml: general mouse list
ma: mouse citation
<<iterative>>
Secondary Triage
Is article relevant to Phenotype curation?
no
discard article
yes
add article to phenotype curation list
pcl: phenotype article list
<<iterative>>
Phenotype curation
obtain full text / PDF
fta: full text article
mpc: mouse phenotype citation
skim article, examine figures, identify relevant passages
rae : relevant article excerpts
identify all alleles in paper and link to relevant excerpts
al : allele list
ai: allele instance
<<iterative>>
Exhaustive curation of each allele
find passage in text
list associated phenotypes
pl : phenotype list
pi: phenotype instance
<<iterative>>
Phenotype curation given an allele
Is phenotype in ontology?
no
Add new entry to phenotype ontology
yes
Is phenotype representation OK?
no
Edit phenotype ontology
yes
poe : phenotype ontology entry
create new database instance with docID, allele, poe
dbi : database instance
dbil: database instance list

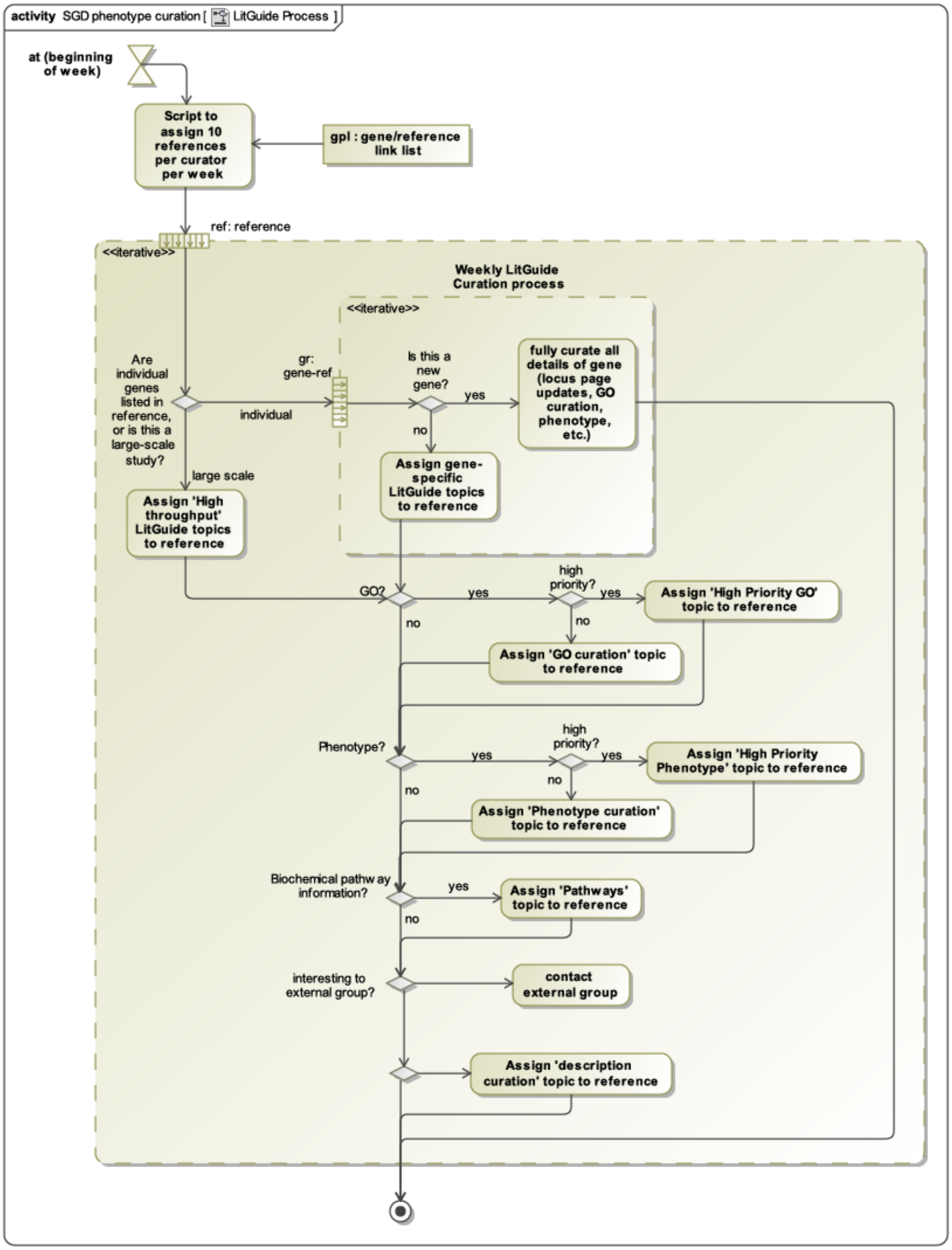
Saccharomyces Genome Database

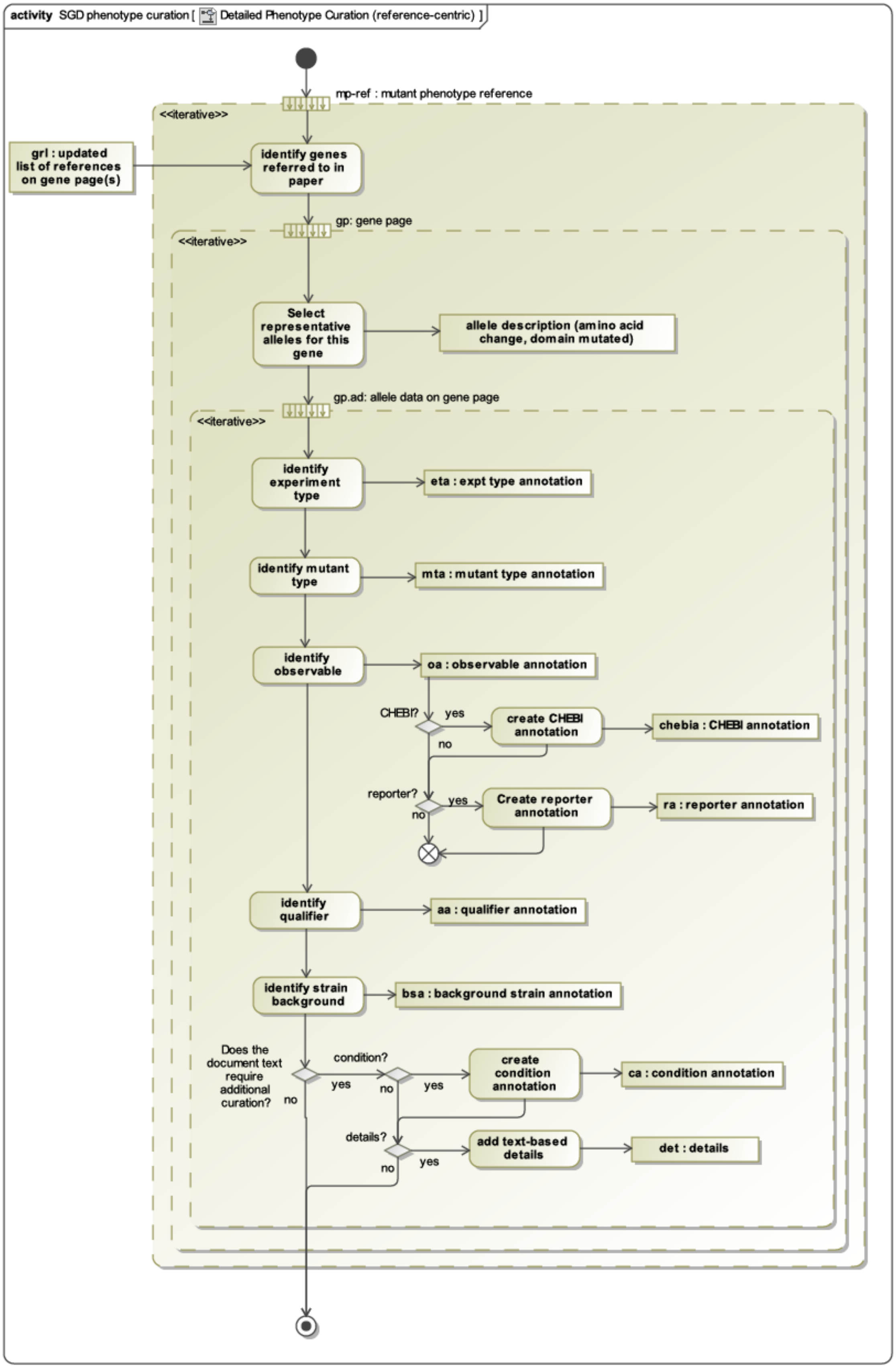
This is a system that processes the literature at 3 stages.

- Automated scripts that execute the triage task of identifying papers of interest
- A specialized process called 'LitGuide' that involves annotators sorting through papers marked for curation and labeling them for later processing
- Detailed Phenotype Curation

There are also circumstances where a specific paper is marked for 'high priority curation'. We have not represented this process here.

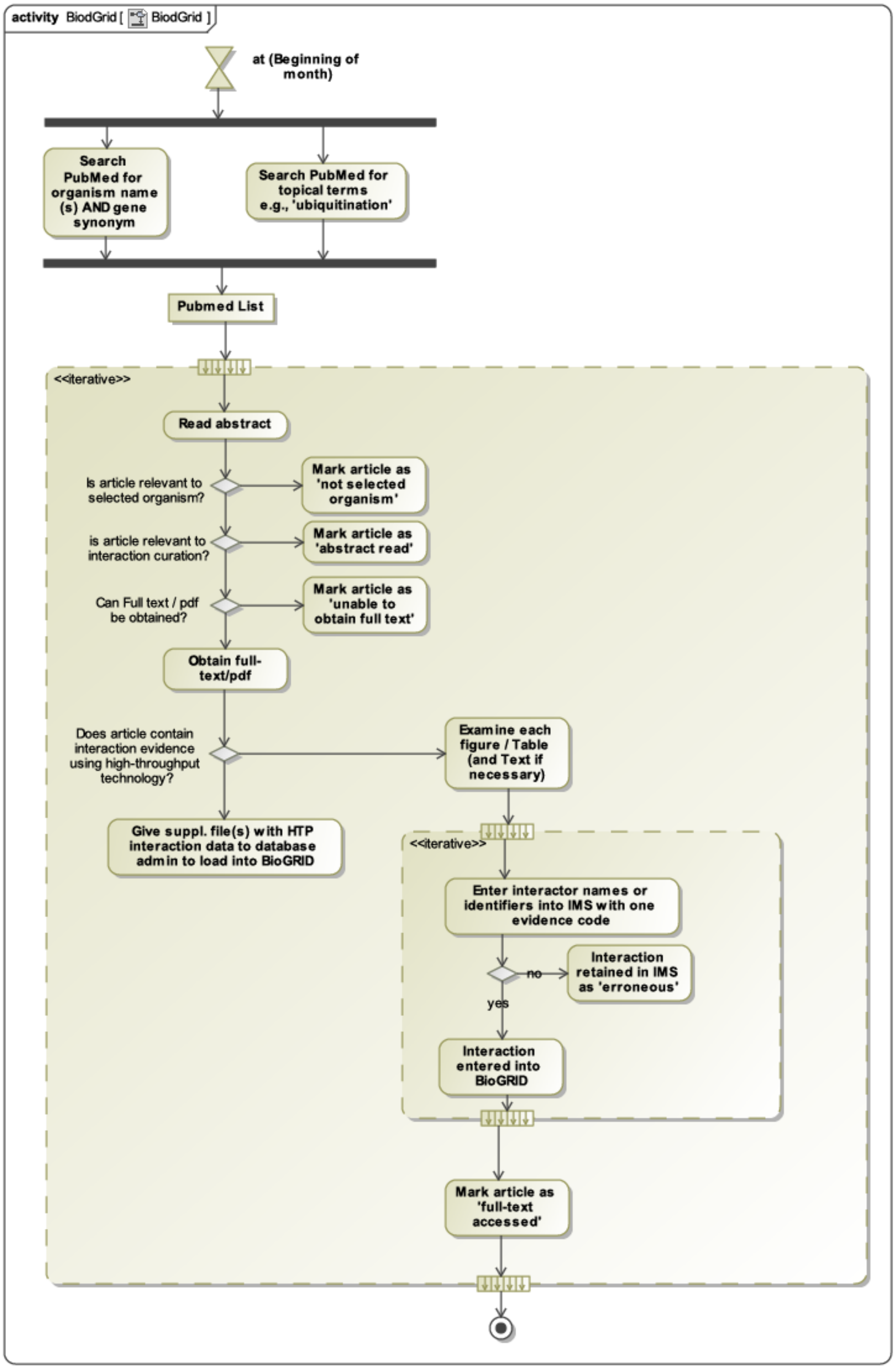

activity SGD phenotype curation [Automatic scripts + screeners]
at (Beginning of month)
run medline script: 'yeast' or 'cerevisiae'
reference list
<<iterative>>
Is this relevant to SGD?
no
yes
brl: bad reference list
assign gene names to reference
cgrl: curated gene + reference list
medline search for gene name patterns + 'cerevisiae'
automatically extract gene names
tgrl : tentative gene/reference link list
reference associated with lit. guide page of gene(s), with status 'not yet curated'
grl : updated list of references on gene page(s)


activity SGD phenotype curation [LitGuide Process]
at (beginning of week)
Script to assign 10 references per curator per week
gpl : gene/reference link list
ref: reference
<<iterative>>
Weekly LitGuide Curation process
<<iterative>>
Are individual genes listed in reference, or is this a large-scale study?
gr: gene-ref
individual
Is this a new gene?
yes
fully curate all details of gene (locus page updates, GO curation, phenotype, etc.)
no
Assign gene-specific LitGuide topics to reference
large scale
Assign 'High throughput' LitGuide topics to reference
GO?
yes
high priority?
yes
Assign 'High Priority GO' topic to reference
no
no
Assign 'GO curation' topic to reference
Phenotype?
yes
high priority?
yes
Assign 'High Priority Phenotype' topic to reference
no
no
Assign 'Phenotype curation' topic to reference
Biochemical pathw ay information?
yes
Assign 'Pathways' topic to reference
no
interesting to external group?
contact external group
Assign 'description curation' topic to reference


activity SGD phenotype curation [Detailed Phenotype Curation (reference-centric)]
mp-ref : mutant phenotype reference
<<iterative>>
grl : updated list of references on gene page(s)
identify genes referred to in paper
gp: gene page
<<iterative>>
Select representative alleles for this gene
allele description (amino acid change, domain mutated)
gp.ad: allele data on gene page
<<iterative>>
identify experiment type
eta : expt type annotation
identify mutant type
mta : mutant type annotation
identify observable
oa : observable annotation
CHEBI?
yes
no
create CHEBI annotation
chebia : CHEBI annotation
reporter?
yes
no
Create reporter annotation
ra : reporter annotation
identify qualifier
aa : qualifier annotation
identify strain background
bsa : background strain annotation
Does the document text require additional curation?
no
yes
condition?
no
yes
create condition annotation
ca : condition annotation
details?
yes
no
add text-based details
det : details

Protein Databases

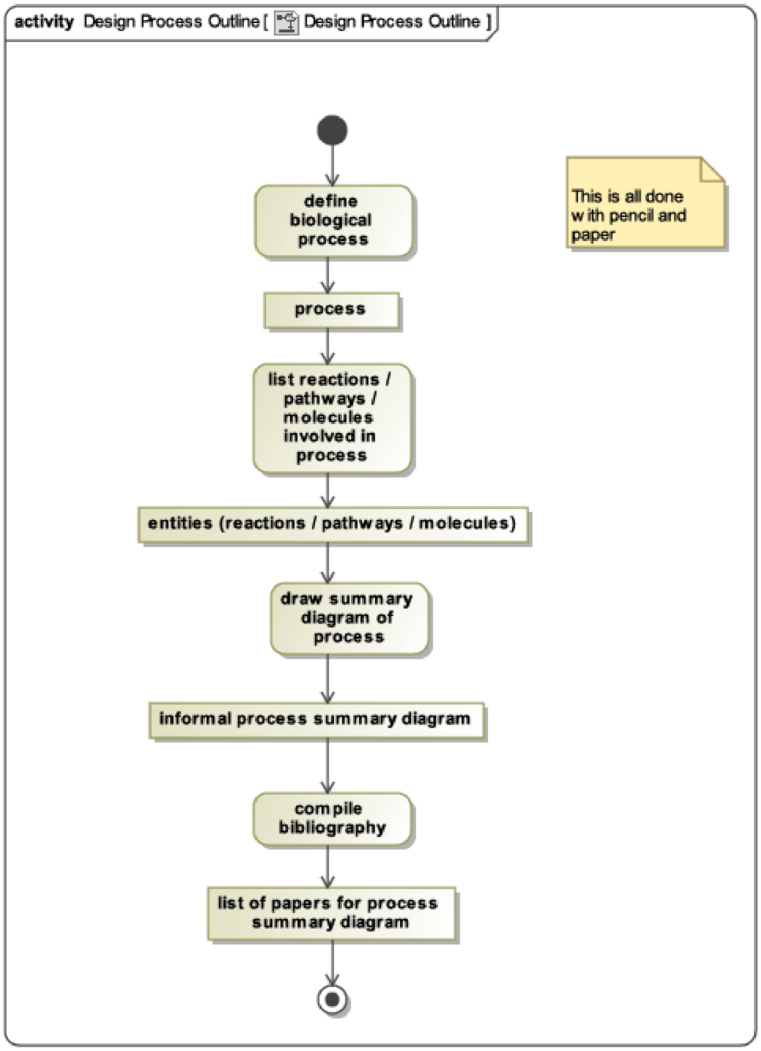
BioGRID

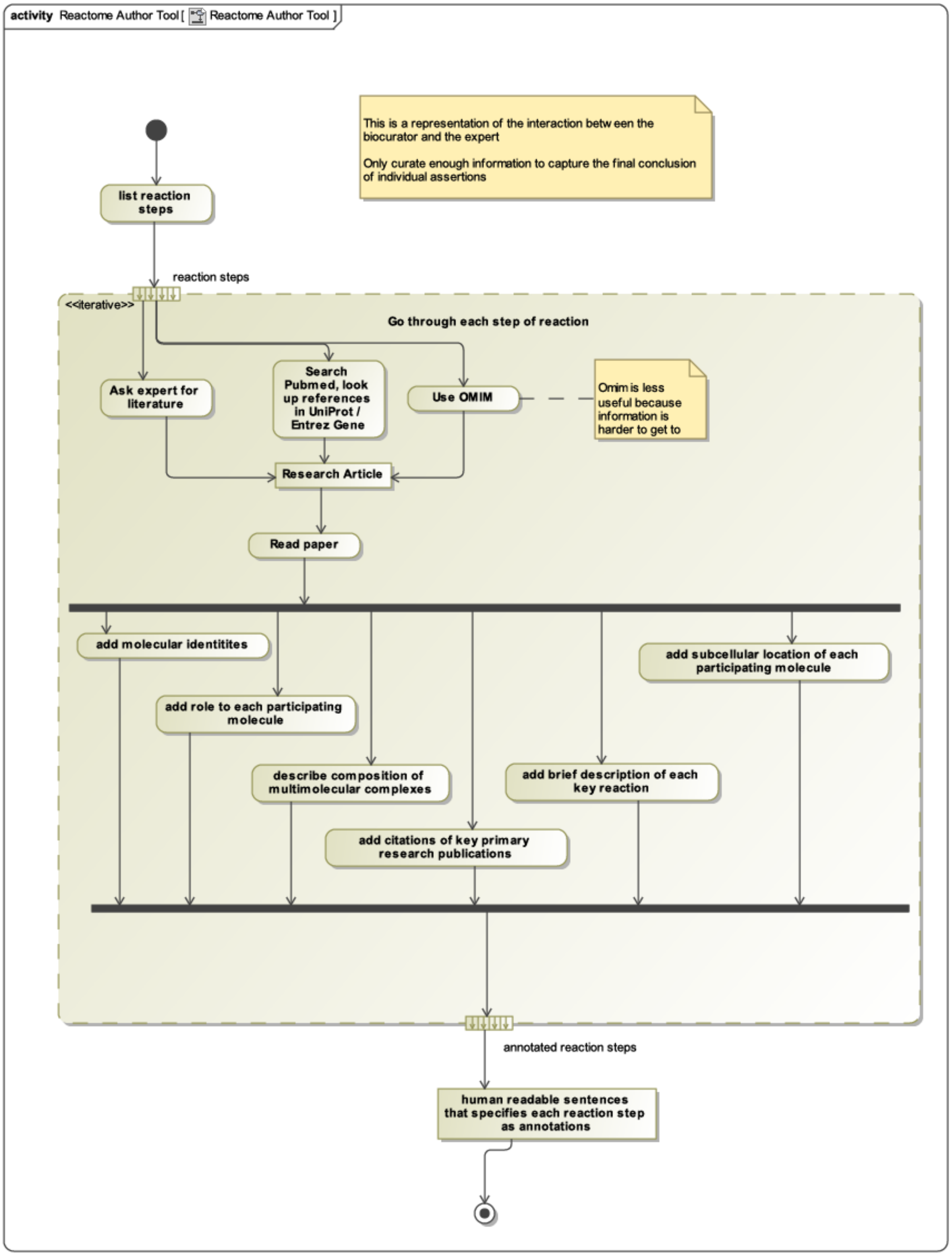

activity BiodGrid [BiodGrid]
at (Beginning of month)
Search PubMed for organism name (s) AND gene synonym
Search PubMed for topical terms e.g., 'ubiquitination'
Pubmed List
<<iterative>>
Read abstract
Is article relevant to selected organism?
Mark article as 'not selected organism'
is article relevant to interaction curation?
Mark article as 'abstract read'
Can Full text / pdf be obtained?
Mark article as 'unable to obtain full text'
Obtain full-text/pdf
Does article contain interaction evidence using high-throughput technology?
Examine each figure / Table (and Text if necessary)
Give suppl. file(s) with HTP interaction data to database admin to load into BioGRID
<<iterative>>
Enter interactor names or identifiers into IMS with one evidence code
no
Interaction retained in IMS as 'erroneous'
yes
Interaction entered into BioGRID
Mark article as 'full-text accessed'

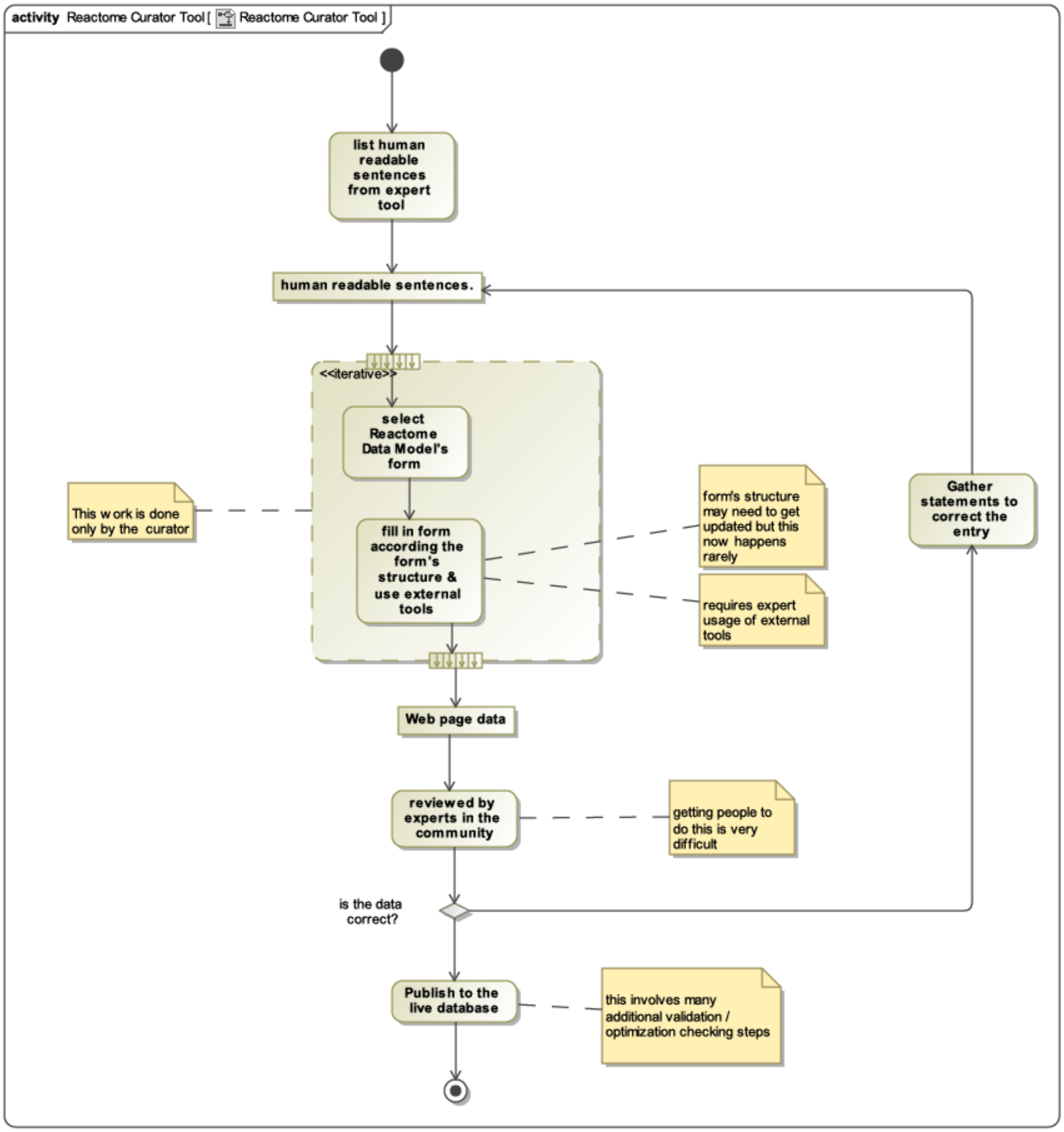
Gallus Reactome

This process is divided into three sections and represents an interaction between an subject matter expert and a biocurator.

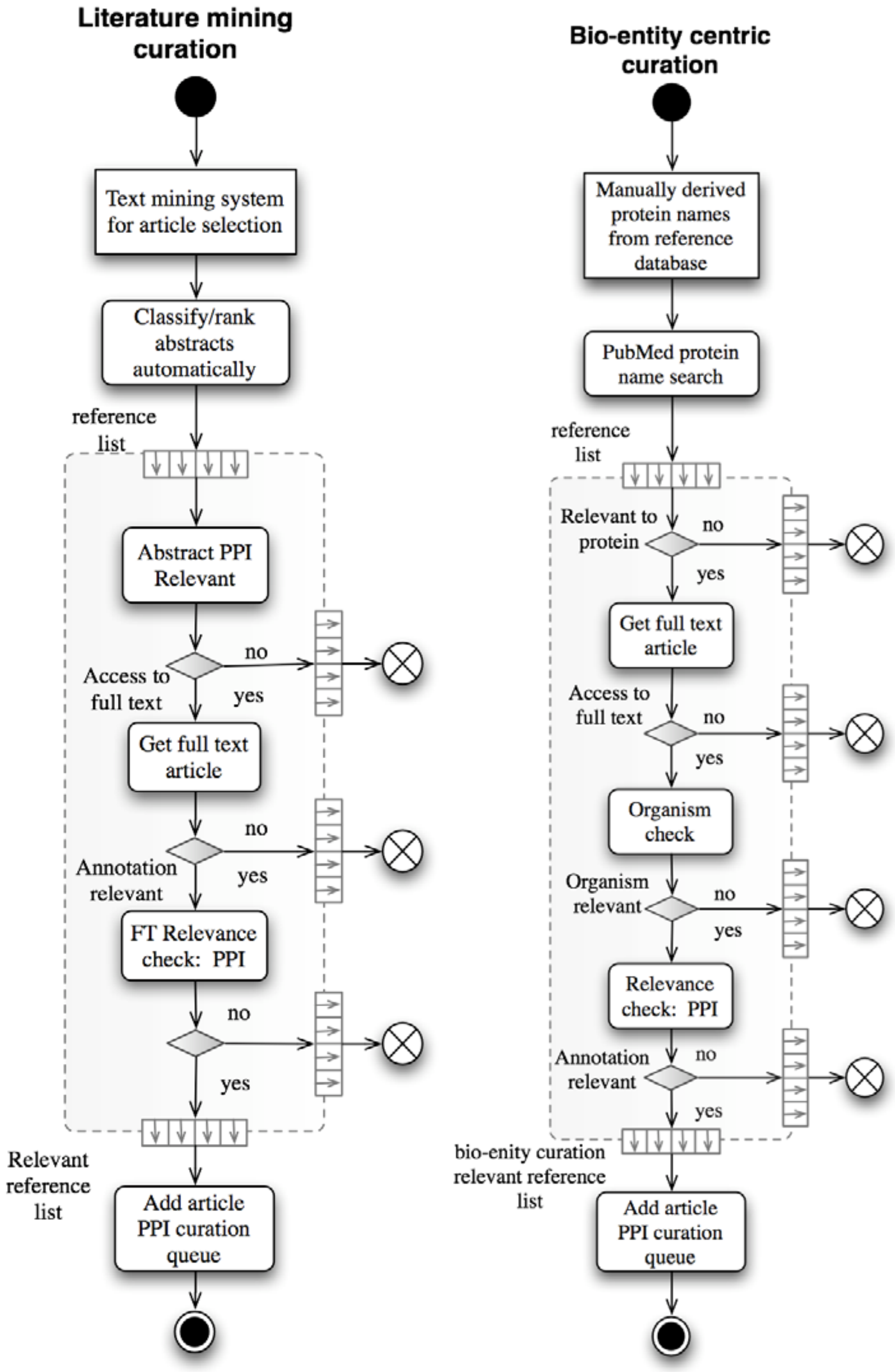
- Design Process Outline : the subject matter expert sketches out a biological process
- Reactome Author Tool : the expert and curator discuss the process of understanding the biological process sketch and build statements that are well understood by the curator
- Reactome Curator Tool : the curator processes each statement by entering them into the standard data entry forms within the Reactome curation system.

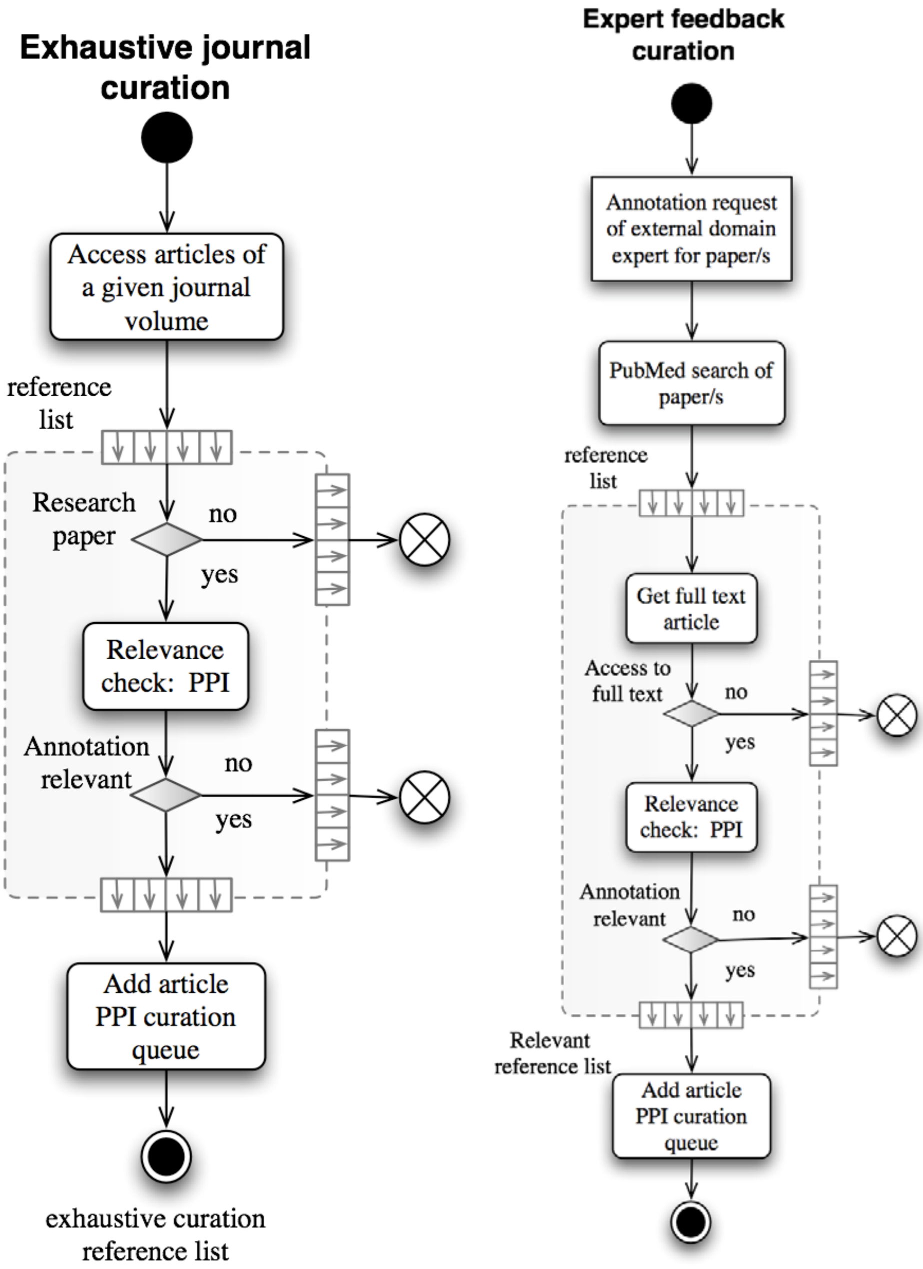

activity Design Process Outline [Design Process Outline]
define biological process
process
list reactions / pathways / molecules involved in process
entities (reactions / pathways / molecules)
draw summary diagram of process
informal process summary diagram
compile bibliography
list of papers for process summary diagram
This is all done with pencil and paper

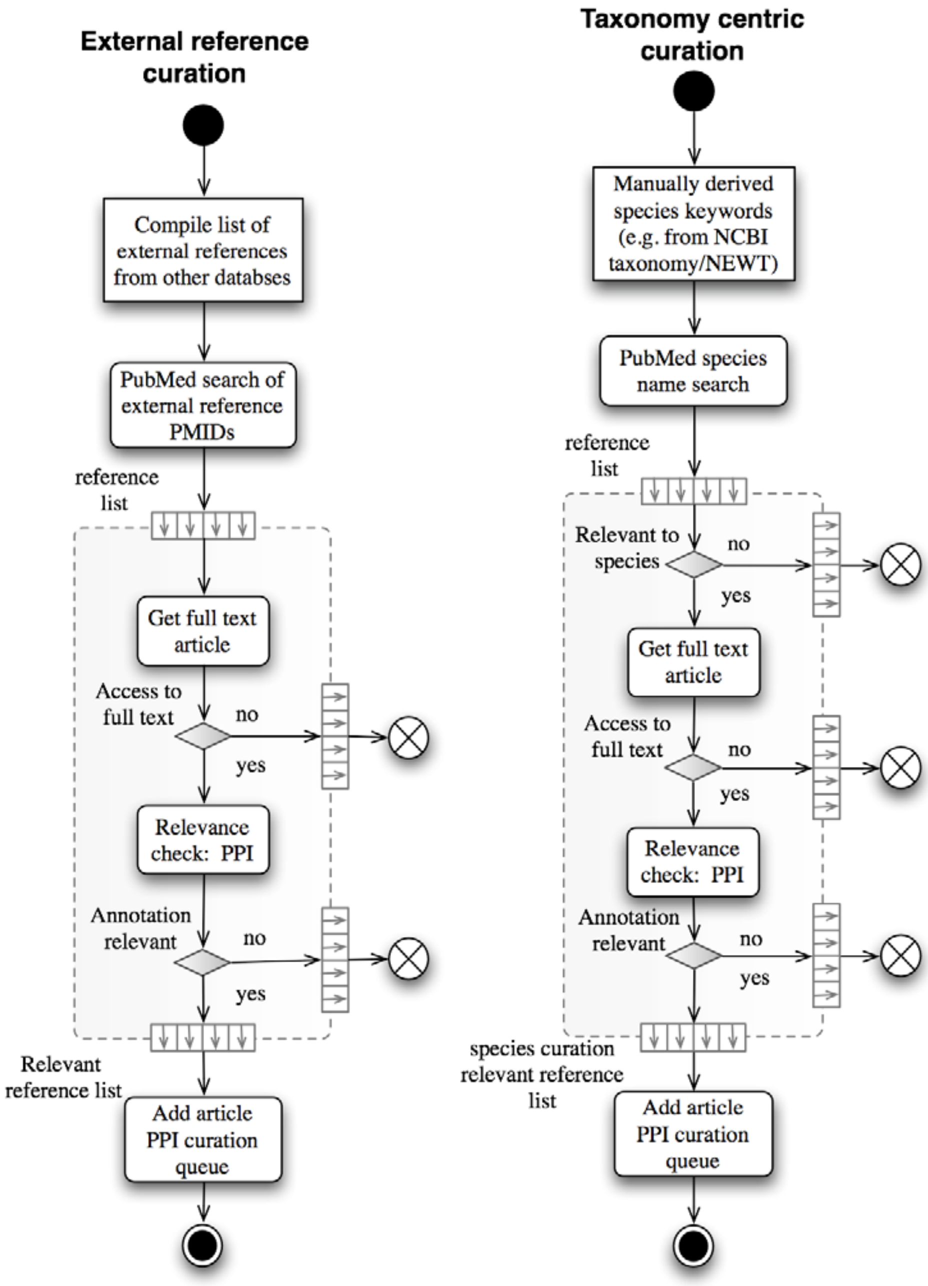

activity Reactome Author Tool [Reactome Author Tool]
This is a representation of the interaction betw een the biocurator and the expert
Only curate enough information to capture the final conclusion of individual assertions
list reaction steps
reaction steps
<<iterative>>
Go through each step of reaction
Ask expert for literature
Search Pubmed, look up references in UniProt / Entrez Gene
Use OMIM
Omim is less useful because information is harder to get to
Research Article
Read paper
add molecular identitites
add role to each participating molecule
describe composition of multimolecular complexes
add citations of key primary research publications
add brief description of each key reaction
add subcellular location of each participating molecule
annotated reaction steps
human readable sentences that specifies each reaction step as annotations


activity Reactome Curator Tool [Reactome Curator Tool]
list human readable sentences from expert tool
human readable sentences.
<<iterative>>
select Reactome Data Model's form
fill in form according the form's structure & use external tools
This w ork is done only by the curator
form's structure may need to get updated but this now happens rarely
requires expert usage of external tools
Gather statements to correct the entry
Web page data
reviewed by experts in the community
getting people to do this is very difficult
is the data correct?
Publish to the live database
this involves many additional validation / optimization checking steps

PPI


Literature mining curation
Text mining system for article selection
Classify/rank abstracts automatically
reference list
Abstract PPI Relevant
Access to full text
no
yes
Get full text article
Annotation relevant
no
yes
FT Relevance check: PPI
no
yes
Relevant reference list
Add article PPI curation queue
Bio-entity centric curation
Manually derived protein names from reference database
PubMed protein name search
reference list
Relevant to protein
no
yes
Get full text article
Access to full text
no
yes
Organism check
Organism relevant
no
yes
Relevance check: PPI
Annotation relevant
no
yes
bio-enity curation relevant reference list
Add article PPI curation queue

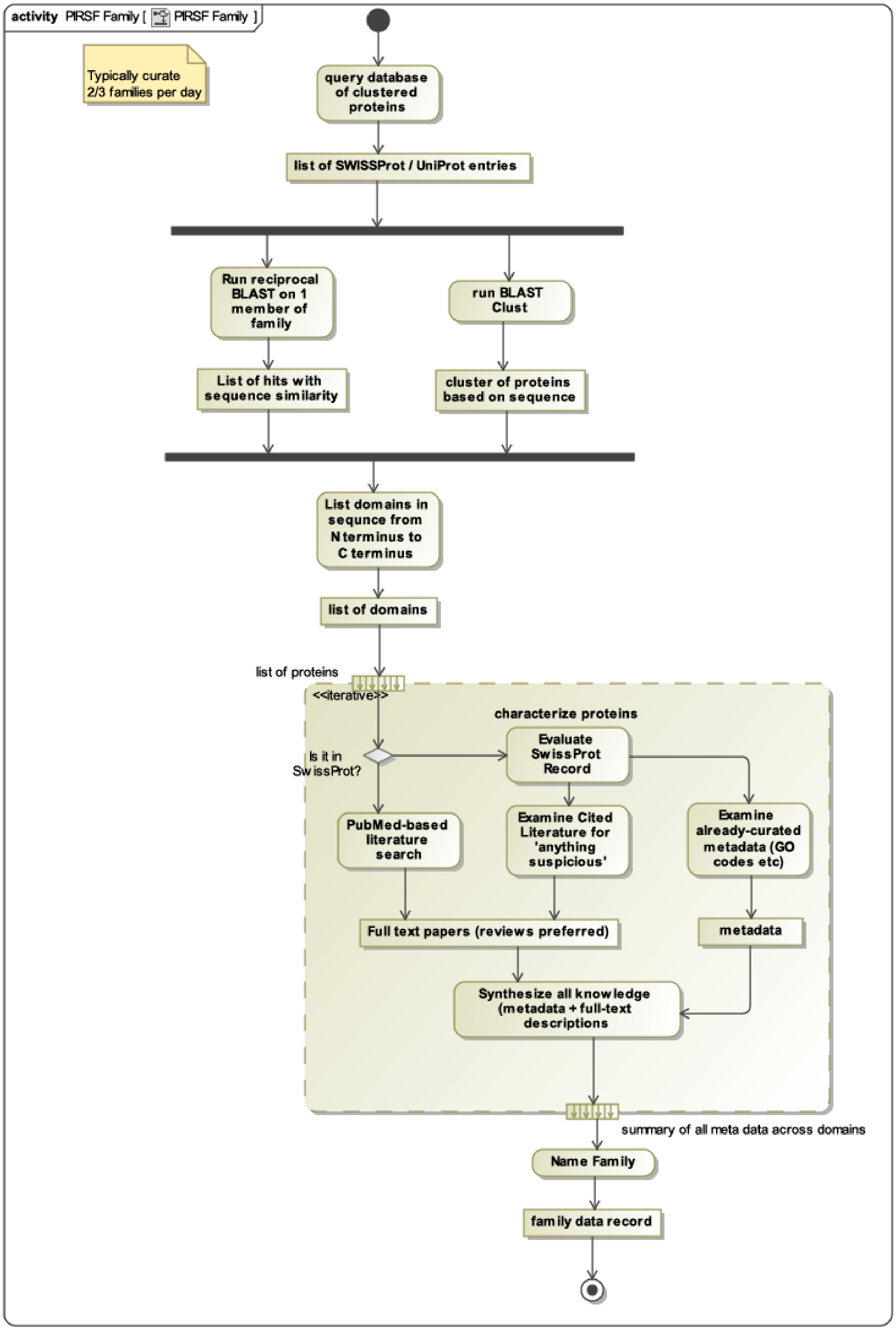

Exhaustive journal curation
Access articles of a given journal volume
reference list
Research paper
no
yes
Relevance check: PPI
Annotation relevant
no
yes
Add article PPI curation queue
exhaustive curation reference list
Expert feedback curation
Annotation request of external domain expert for paper/s
PubMed search of paper/s
reference list
Get full text article
Access to full text
no
yes
Relevance check: PPI
Annotation relevant
no
yes
Relevant reference list
Add article PPI curation queue

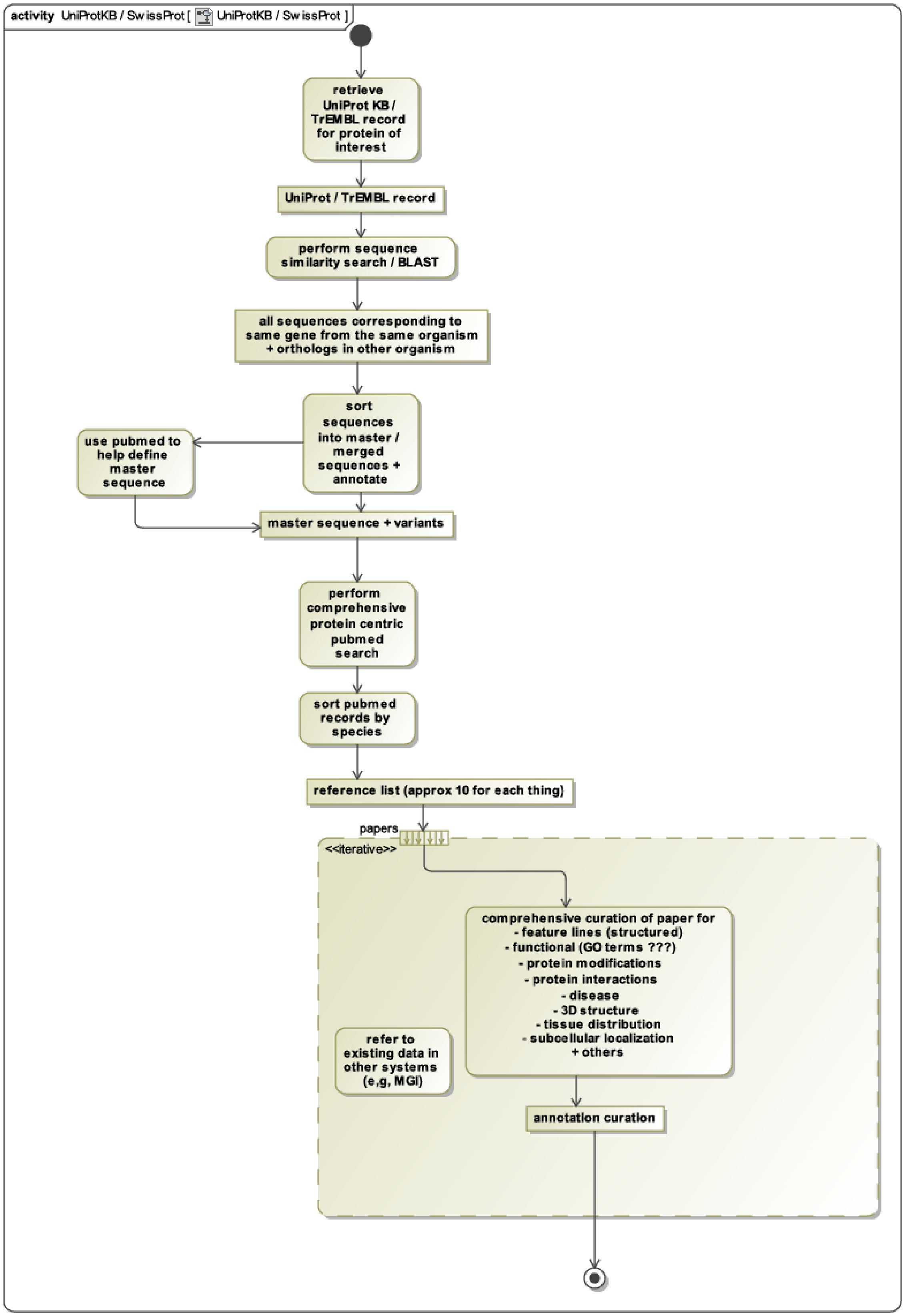

External reference curation
Compile list of external references from other databses
PubMed search of external reference PMIDs
reference list
Get full text article
Access to full text
no
yes
Relevance check: PPI
Annotation relevant
no
yes
Relevant reference list
Add article PPI curation queue
Taxonomy centric curation
Manually derived species keywords (e.g. from NCBI taxonomy/NEWT)
PubMed species name search
reference list
Relevant to species
no
yes
Get full text article
Access to full text
no
yes
Relevance check: PPI
Annotation relevant
no
yes
species curation relevant reference list
Add article PPI curation queue

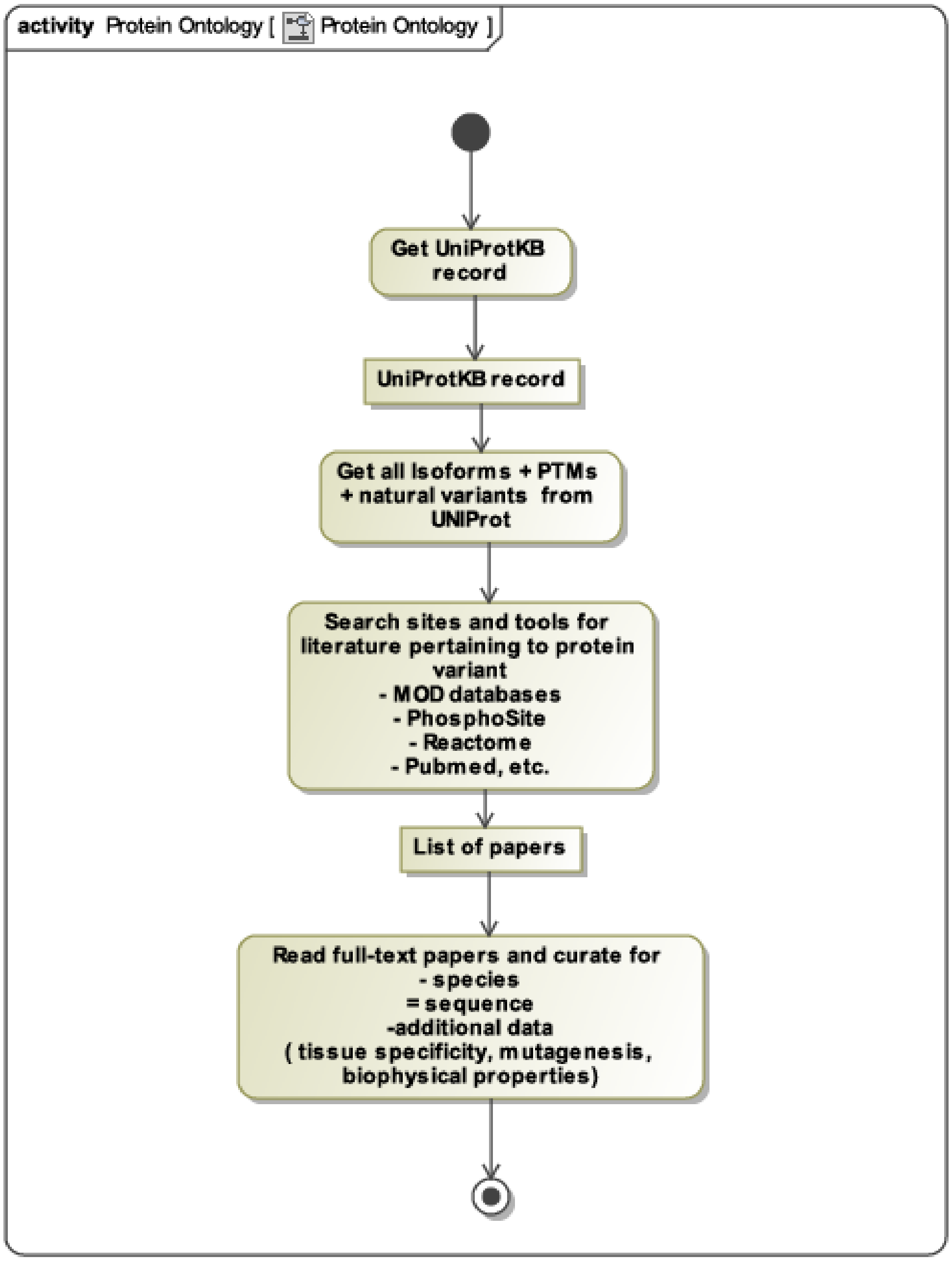
UniProtKB / SwissProt

This curation process involves three elements

- PIRSF Family: This is the process used by biocurators to name protein families
- UniProtKB / SwissProt: This is the process used to curate information into the UniProtKB / SwissProt database
- Protein Ontology

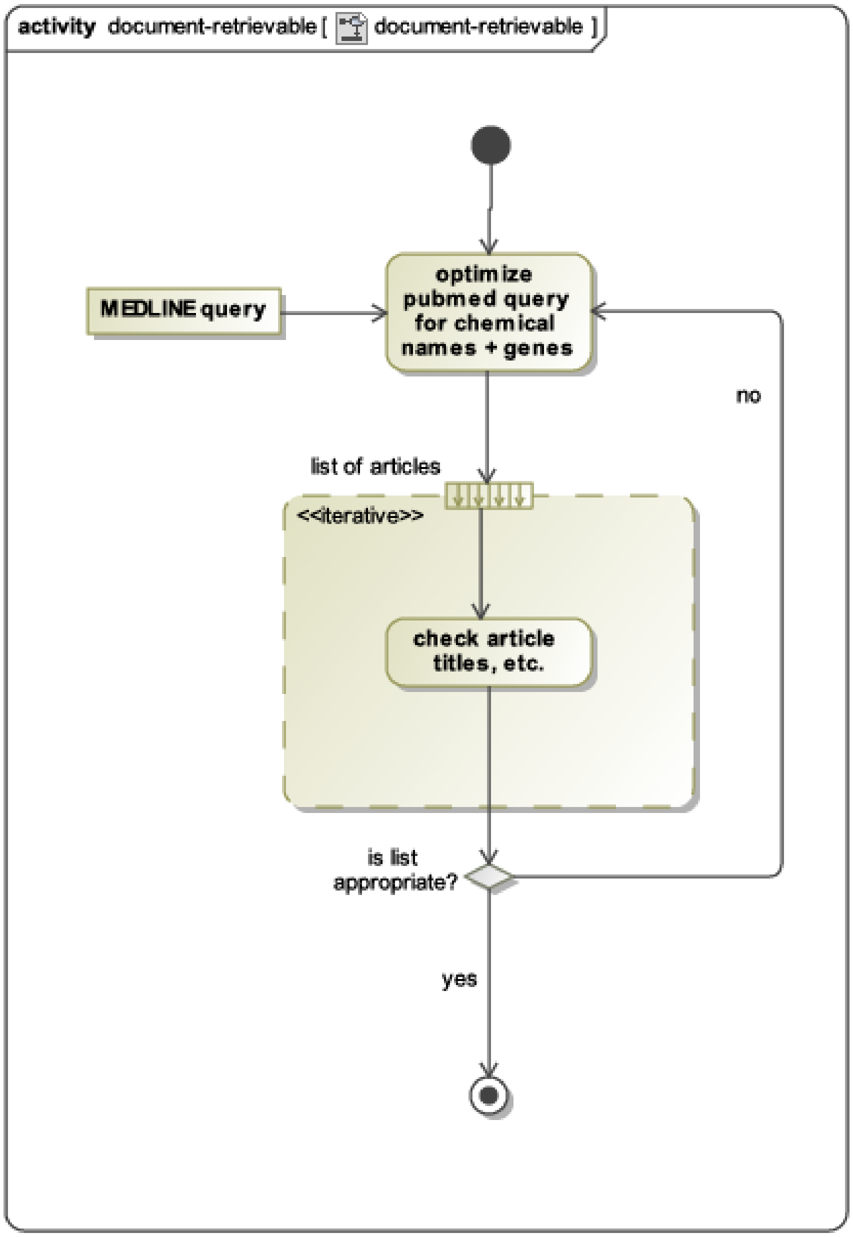

activity PIRSF Family [PIRSF Family]
Typically curate 2/3 families per day
query database of clustered proteins
list of SWISSProt / UniProt entries
Run reciprocal BLAST on 1 member of family
run BLAST Clust
List of hits with sequence similarity
cluster of proteins based on sequence
List domains in sequnce from N terminus to C terminus
list of domains
list of proteins
<<iterative>>
characterize proteins
Is it in SwissProt?
Evaluate SwissProt Record
PubMed-based literature search
Examine Cited Literature for 'anything suspicious'
Examine already-curated metadata (GO codes etc)
Full text papers (reviews preferred)
metadata
Synthesize all knowledge (metadata + full-text descriptions
summary of all meta data across domains
Name Family
family data record

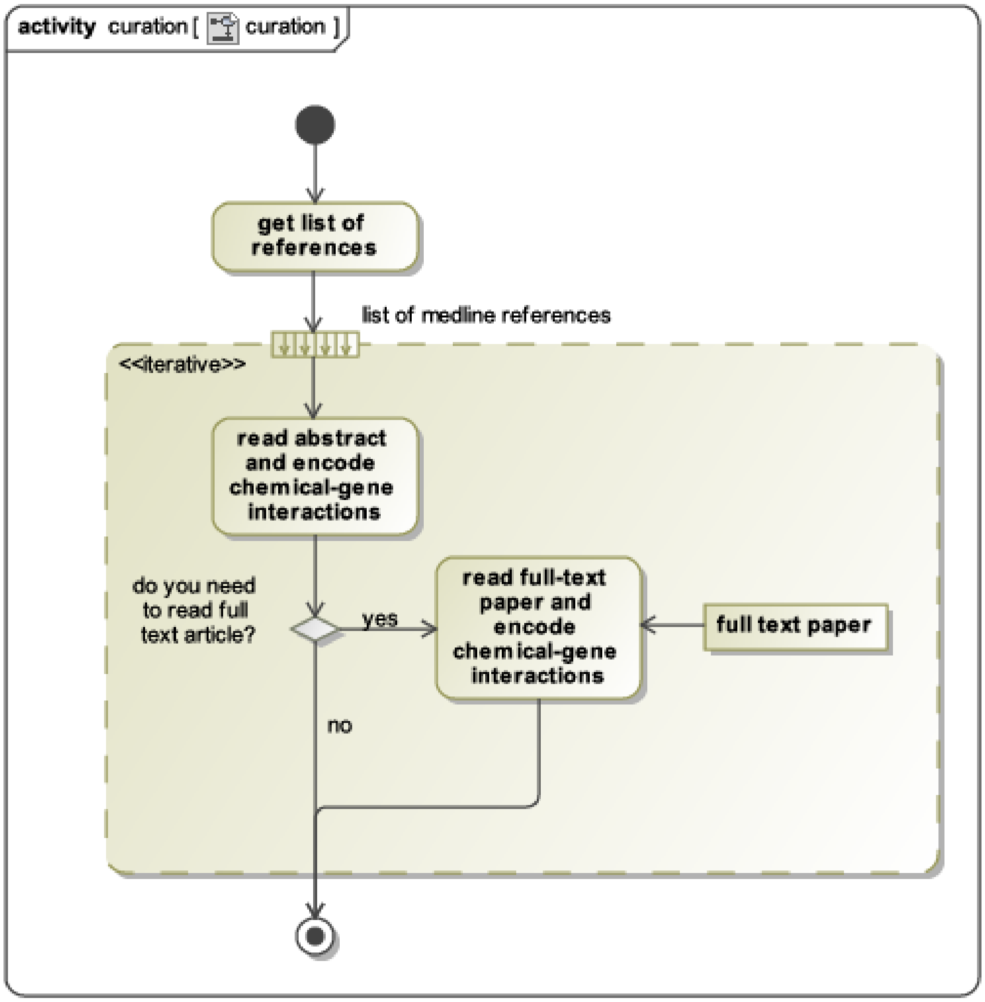

activity UniProtKB / SwissProt [UniProtKB / SwissProt]
retrieve UniProt KB / TrEMBL record for protein of interest
UniProt / TrEMBL record
perform sequence similarity search / BLAST
all sequences corresponding to same gene from the same organism + orthologs in other organism
sort sequences into master / merged sequences + annotate
use pubmed to help define master sequence
master sequence + variants
perform comprehensive protein centric pubmed search
sort pubmed records by species
reference list (approx 10 for each thing)
papers
<<iterative>>
comprehensive curation of paper for
- feature lines (structured)
- functional (GO terms ???)
- protein modifications
- protein interactions
- disease
- 3D structure
- tissue distribution
- subcellular localization
+ others
refer to existing data in other systems (e,g, MGI)
annotation curation


activity Protein Ontology [Protein Ontology]
Get UniProtKB record
UniProtKB record
Get all Isoforms + PTMs + natural variants from UNIProt
Search sites and tools for literature pertaining to protein variant
- MOD databases
- PhosphoSite
- Reactome
- Pubmed, etc.
List of papers
Read full-text papers and curate for
- species
= sequence
-additional data
(tissue specificity, mutagenesis, biophysical properties)

Other

Comparative Toxicogenomics Database (CTD)


activity document-retrievable [document-retrievable]
MEDLINE query
optimize pubmed query for chemical names + genes
no
list of articles
<<iterative>>
check article titles, etc.
is list appropriate?
yes


activity curation [curation]
get list of references
list of medline references
<<iterative>>
read abstract and encode chemical-gene interactions
do you need to read full text article?
yes
no
read full-text paper and encode chemical-gene interactions
full text paper