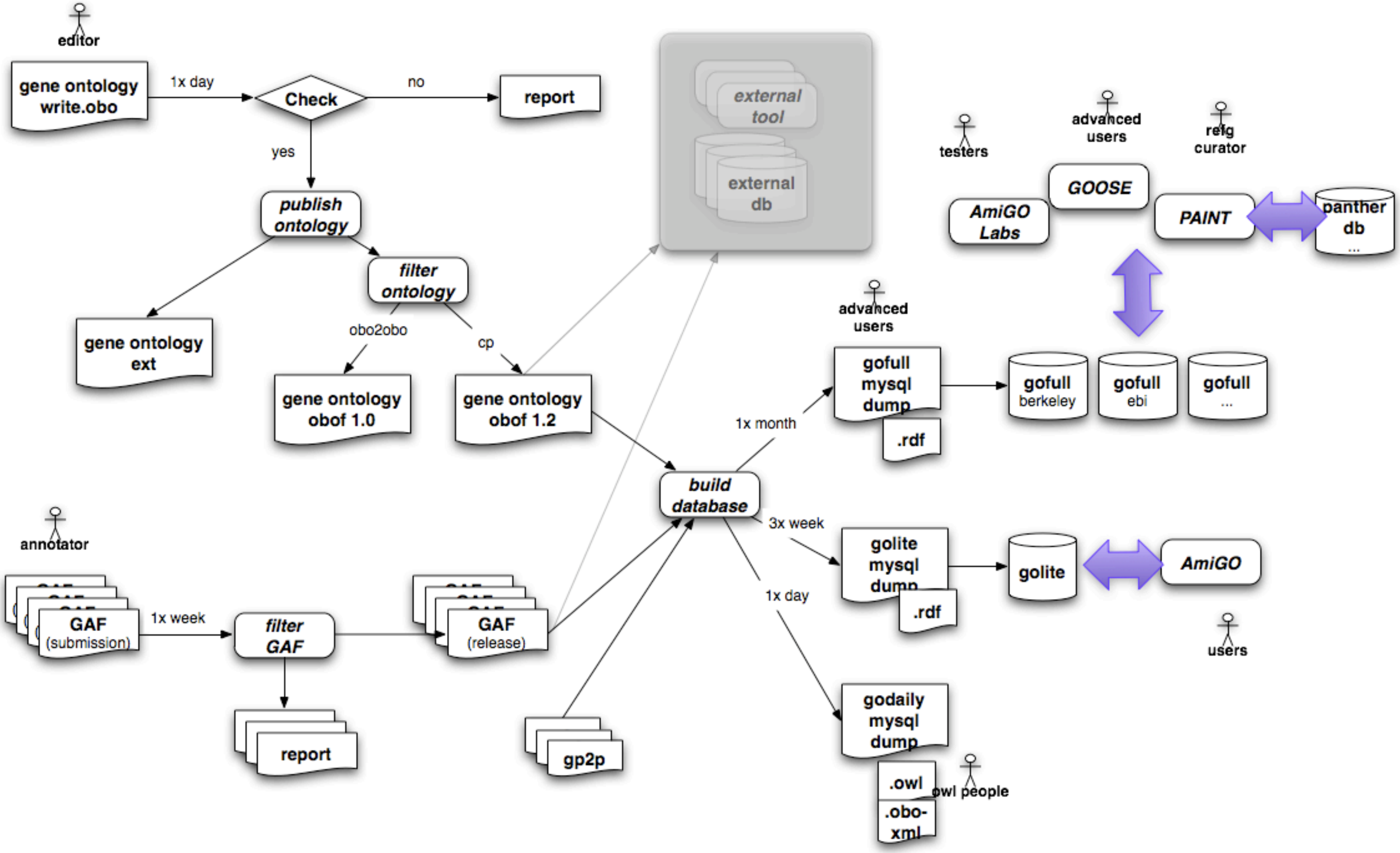
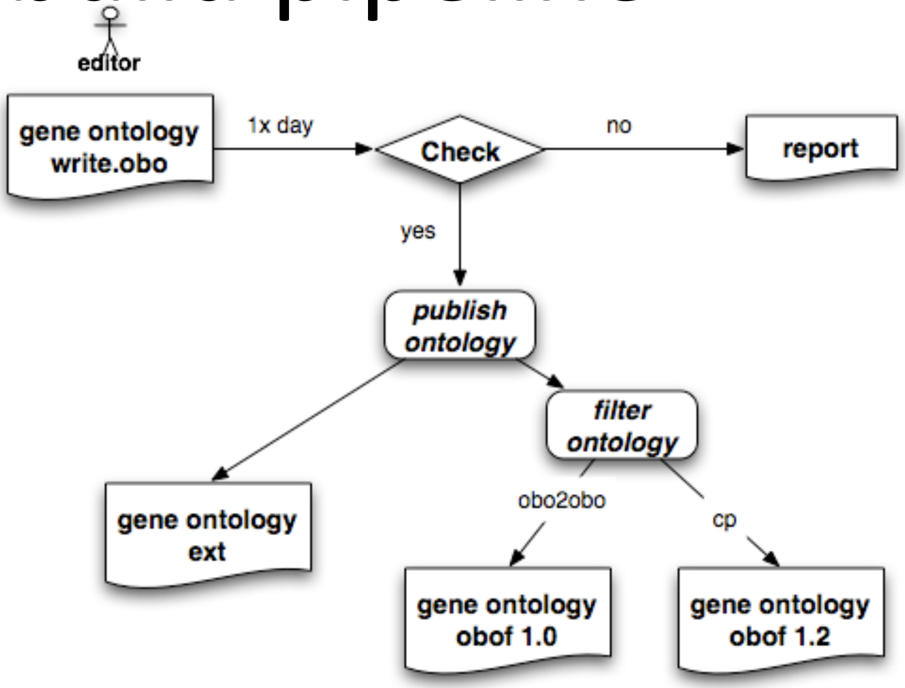
GO Infrastructure

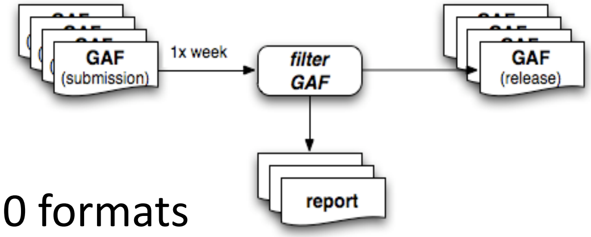

editor
gene ontology write.obo
1x day
Check
no
report
yes
publish ontology
filter ontology
gene ontology ext
obo2obo
cp
gene ontology obof 1.0
gene ontology obof 1.2
external tool
external db
testers
advanced users
refg curator
GOOSE
AmiGO Labs
PAINT
panther db
...
advanced users
gofull mysql dump
.rdf
gofull berkeley
gofull ebi
gofull ...
1x month
build database
3x week
golite mysql dump
.rdf
golite
AmiGO
users
1x day
godaily mysql dump
.owl
.obo-xml
owl people
annotator
GAF (submission)
1x week
filter GAF
report
GAF (release)
gp2p

Ontology build pipeline



- Checks
  - ascii-only
  - namespace
  - is_a-complete
  - disjoint violation
  - duplicate names
- Filtering (not go_ext)
  - inter-ontology links
  - intra-MF regulates
  - obof1.3 tags (created_by, creation_date) [*]
  - **proposed: adds data-version tag (post-sept 2009)**
  - intersection_of
- Less will be filtered in future

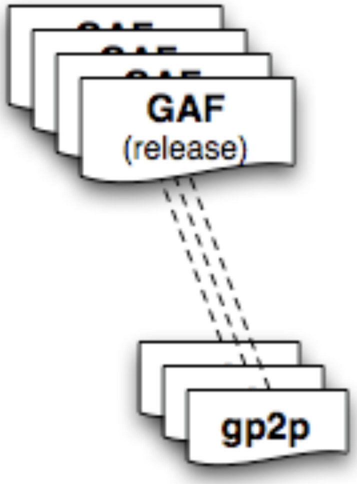
GAF publishing pipeline



- Incoming submissions may **either** be GAF2.0 **or** GAF1.0 formats
- Conversion to GAF1 takes place prior to publishing
  - Strips last 2 columns
  - Users download GAF1 files
  - source GAF2s are available for advanced users
- MGI first to publish GAF2
  - http://cvsweb.geneontology.org/cgi-bin/cvsweb.cgi/go/gene-associations/submission/gene_association.mgi.gz
  - 87 genes with different isoforms
  - col 16 not filled in yet
  - UniProt to follow
- We need to decide on time for the switch to making GAF2 the primary published format
  - producers can submit **either** GAF1 **or** GAF2
  - GAF1 converted to GAF2 in pipeline
  - Should not affect GAF1 parsers

gp2protein and col 17

- include
  - gene (col2) to generic UniProt ID
  - isoform (col17) to UniProt variant ID (other IDs?)
    - where col17 is e.g. a MOD protein ID



Unannotated genes and GAFs

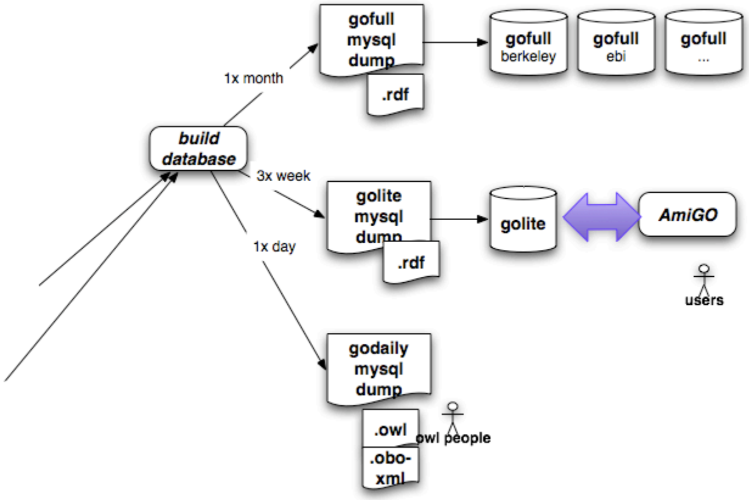
- We currently lack information on unannotated genes
  - GAF only includes lines for
    - annotated genes
    - ND unannotated genes
  - Problematic for statistics etc
- Proposal
  - add lines to GAF for unannotated genes
  - new evidence code

Proposed additions to pipeline

- Taxon constraint checks
  - http://www.geneontology.org/scratch/go-taxon/
  - Filter lines that fail check
    - Alternatively just warn?
  - Instant check on submission?
- GAF inter-ontology inference
  - infer IC BP annotation from MF annotation
    - 47k annotations
    - Extensible to BP->CC
  - http://www.geneontology.org/scratch/gaf-inference/
  - Automatically add or send back to provider?
    - or both
- GAFs from PAINT
  - file per gene family?
  - new submission dir?
- When do we implement?

GO Database Pre-Sept 2009

	frequency	associations	IEAs?	Seqs	Used by	download options
daily	daily	0	-			owl, go-rdf, mysql
lite	3x a week	0.8m	NO	Yes	AmiGO, Paint	go-rdf, mysql, fasta
full	monthly	33m	ALL	No	GOOSE	go-rdf, mysql



GO Database Post-Sept 2009

	frequency	associations	IEAs?	Seqs	
daily	daily	0	-	-	
lite	3x a week§	1.8m	SOME*	Yes	AmiGO, Paint
full	monthly	33m	ALL	No	GOOSE

§ dumps only 1x a week

***As of Sept 2009, includes IEAs for species-centric GAFs only**

- **includes sgd, fb, …, goa_human, goa_chicken**
- **excludes goa_uniprot**

Proposed post-Sep 2009

	frequency	associations	IEAs?	Seqs	
daily	daily	0	-	-	
lite	**1x a week**	1.8m	SOME*	Yes	AmiGO, Paint, GOOSE
~~full~~	~~monthly~~	~~33m~~	~~ALL~~	~~No~~	

Build pipeline synchronicity

- ontology publishing
  - 1x day
- filter-GAFs
  - freq: 1x week
  - Removes
    - annotations to obsoletes
      - Advance warning to GAF providers
    - annotations to alt_ids (results of merges)
      - SHOULD WE DO THIS? no warning given to GAF providers
- golite mysql
  - freq: **3x** week
  - may be out of sync
  - includes annotations to alt_ids?
    - doesn’t seem to be happening so far

GO Database 2010: proposed

	frequency	associations	IEAs?	Seqs	
daily	daily	0	-	-	
lite	**incremental**		SOME*	Yes	
full?	**incremental**	33m	ALL	No	OBO-Edit (read-only), GOOSE, Paint, **AmiGO**

switch to improved incremental updates
load id-mapping table (~1gb)

Infrastructure software changes

- GOBO
  - Underlying perl code is being refactored
    - based on Moose framework
    - Integrated with next version of bioperl
  - Improvements
    - Better handling of ontology rules
    - new inference checks and taxon filtering uses new framework
    - new map2slim and enricher will treat relations properly
    - will be used in newer AmiGO code
  - Downsides
    - harder to install, more dependencies
- Lucene indexing [fast]
  - [demo]

Database schema extensions

- Phylogeny support
  - Phylogenetic tree support
  - Load panther trees into GODB
  - Display in AmiGO [soon]
- Reasoner code integrated into DB
  - correct behavior for relations
  - Choice of whether to include regulation in slimming, enrichment etc

Proposed change to ontology development pipeline

- Sourceforge is inefficient
  - clunky interface
  - lots of individual requests for trivial compositional terms
  - Lag to get ID
- A subset of term requests could be managed far more efficiently
  - regulation
  - part-specific subtypes
- Fill in basic cross-product info in web interface
  - get back instant GO ID
  - label, synonym and definitions generated automatically
  - reasoner places new term(s) correctly in DAG
    - requires XPs to be live
- [DEMO]
  - http://amigo.berkeleybop.org/cgi-bin/amigo/amigo_exp?mode=xp_term_request_client