Using ontologies to make bioassay protocols machine readable

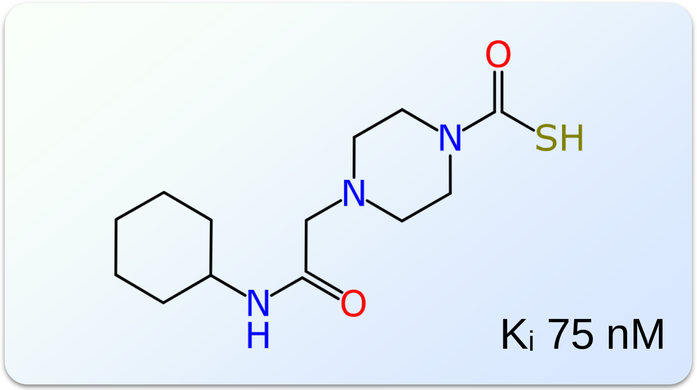
Alex M. Clark

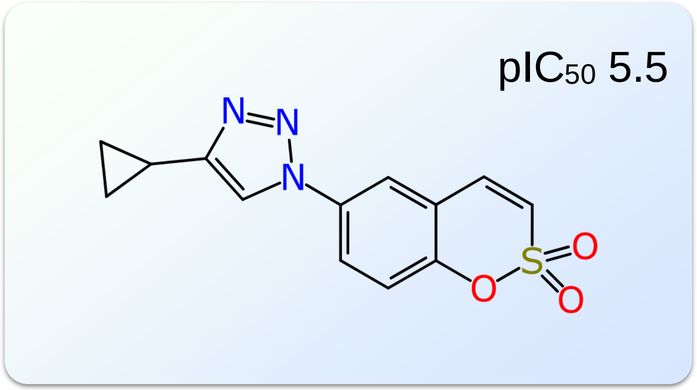
alex@collaborativedrug.com

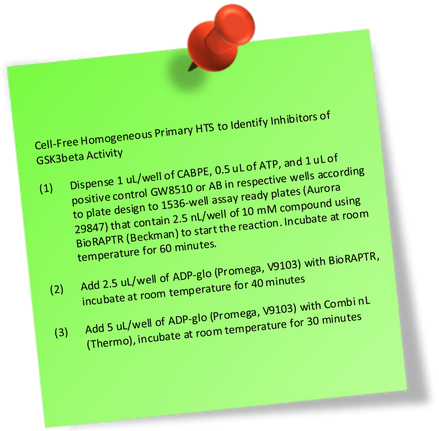
Background

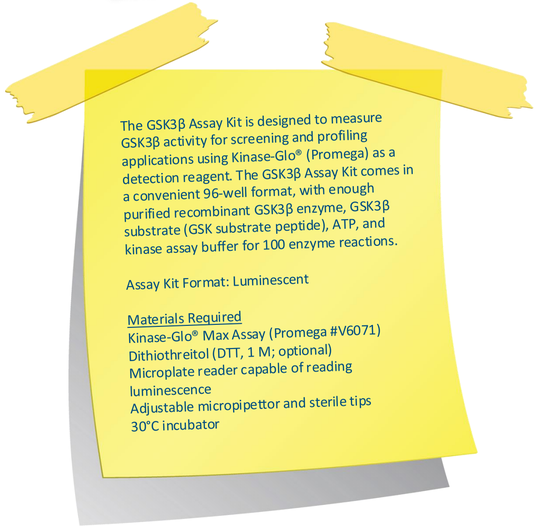
- **chem**informatics and **bio**informatics: machine readable for decades
- ... **assay** informatics under-developed: measurements are rarely FAIR

machine readable









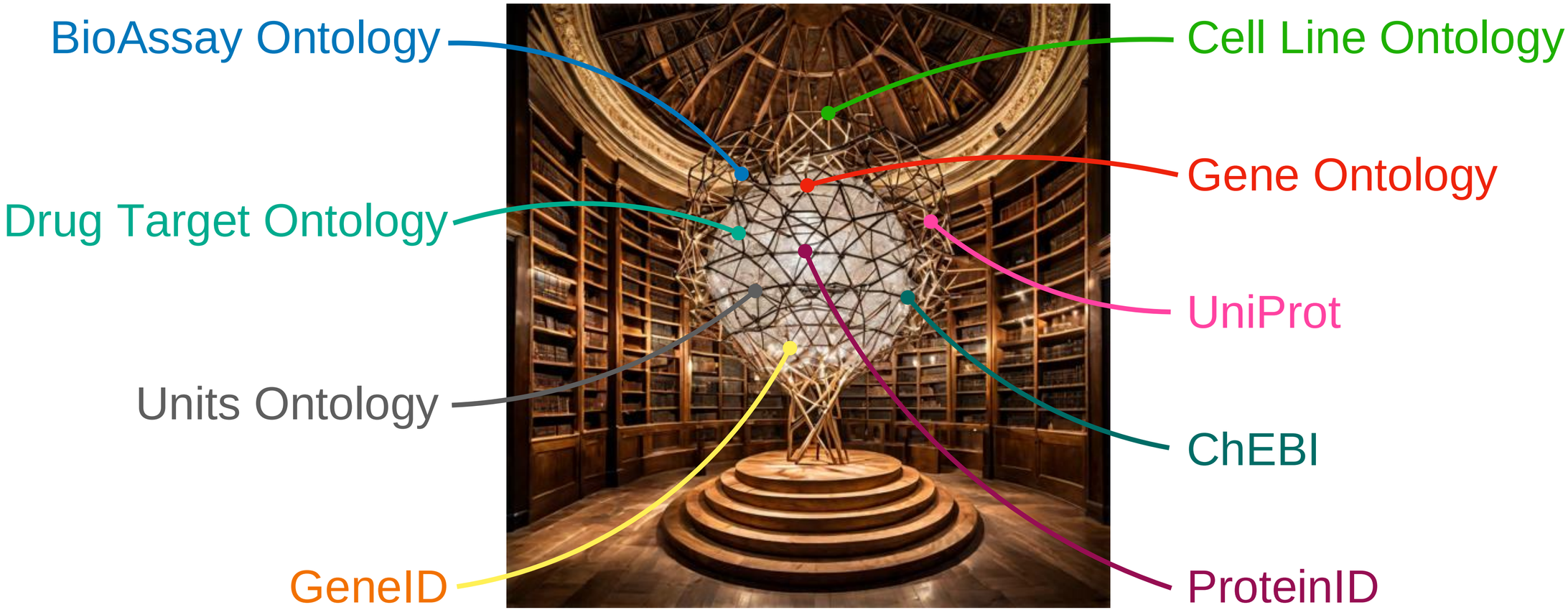
not so much...

Current best practices

- Institutions often have partially systematic assay protocol ELNs
  - e.g. *organism* or *protein target* has well defined dictionary
  - other data mostly made up of
    - **free text**
    - **pick lists**
    - **mnemonic titles**
    - **Word docs**
- Consistency between institutions is rare
- Publications usually terse, inter-referenced or implied

Ontologies to the rescue

- Public, standardized, well defined, but... how to use them?



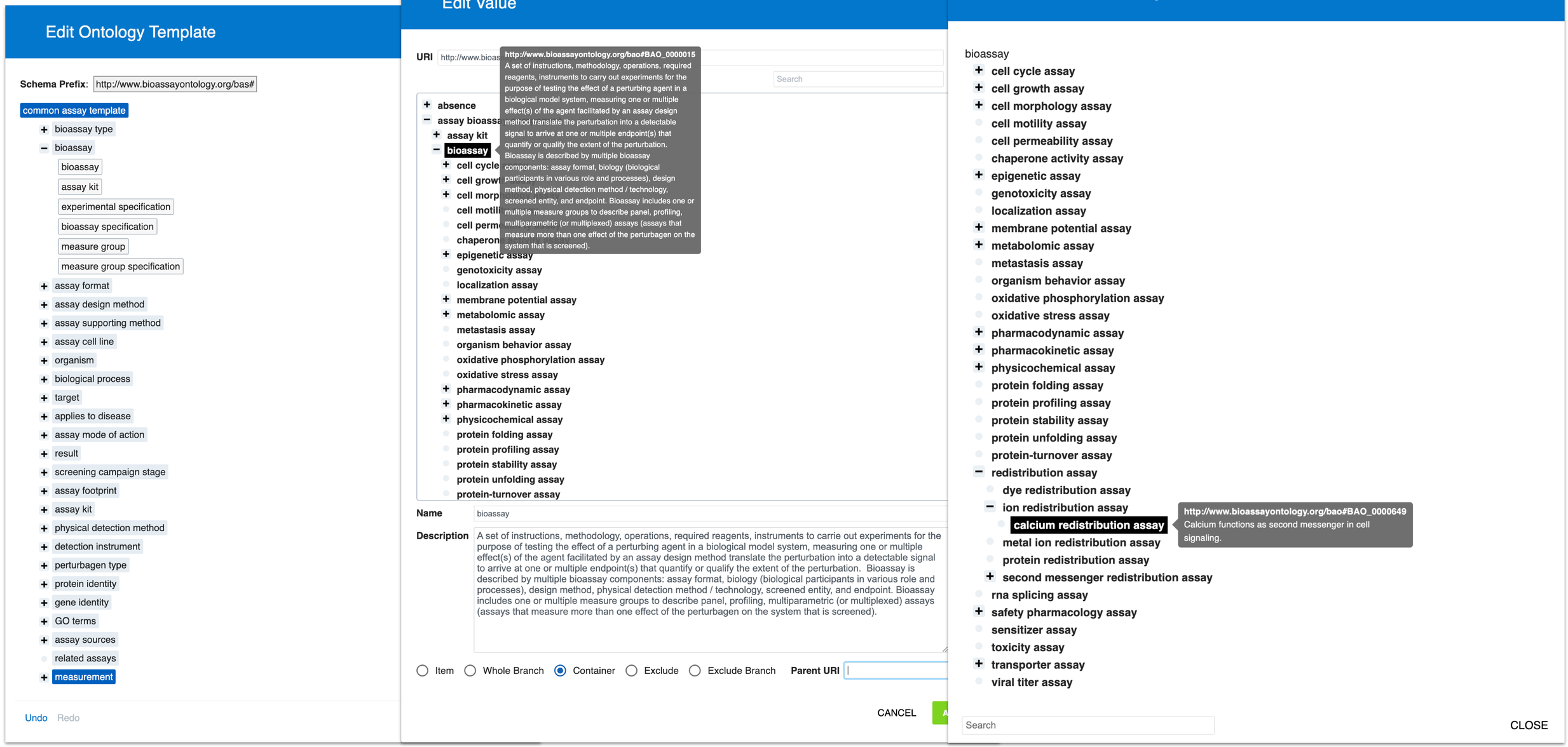
Ontology templates

- Intermediate step:
  - select ontologies relevant to the domain
  - cherry-pick branches and place them into categories
- **Common Assay Template**: most of what you need to know for screening

bioassay type
bioassay
format
design method
supporting method
cell line
organism
biological process
target
applies to disease
mode of action
result
campaign stage
footprint
assay kit
physical detection method
detection instrument
perturbagen type
protein identity
gene identity
GO terms
sources

BioAssay templates for the semantic web
10.7717/peerj-cs.61

Template editor

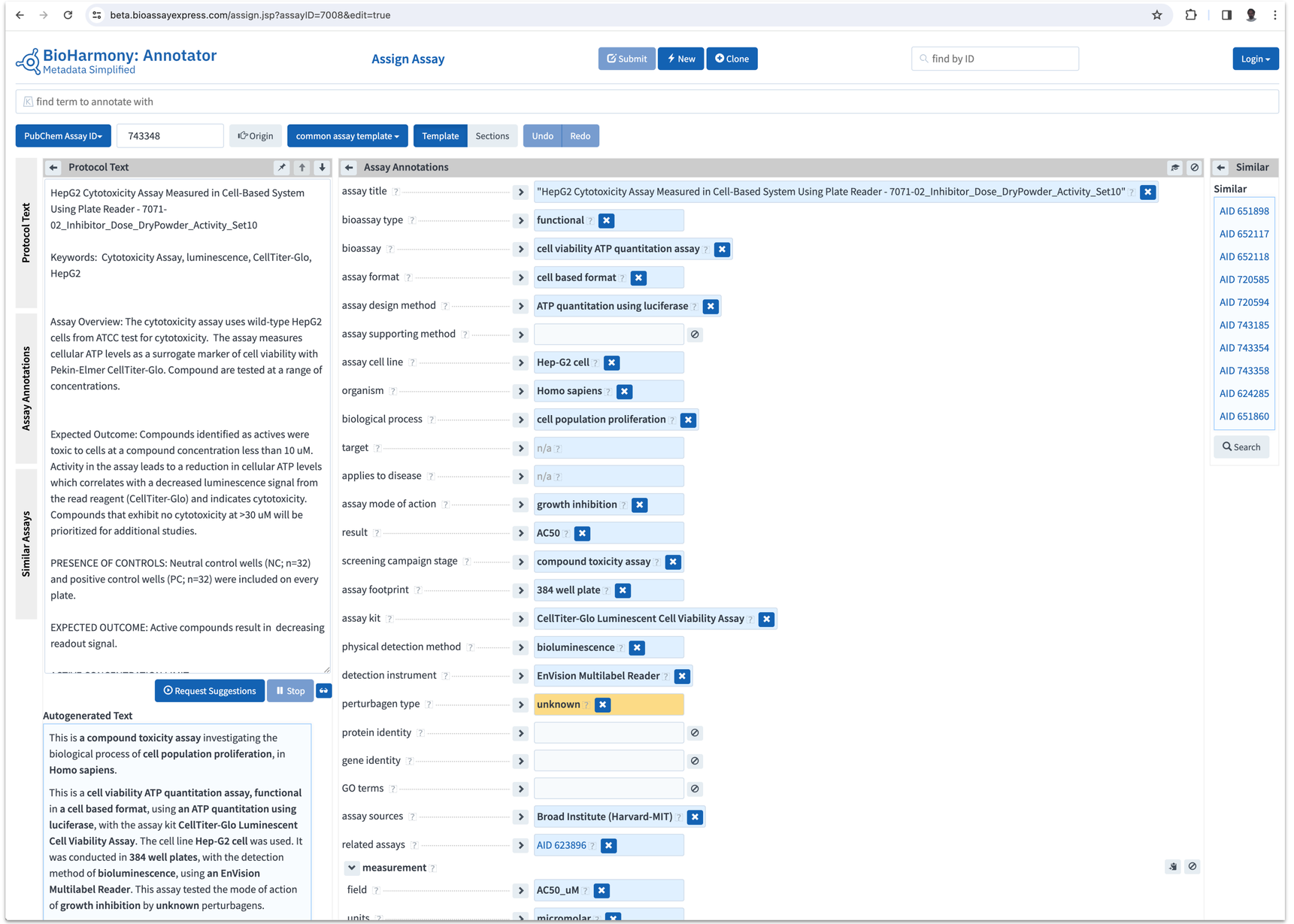


Body of practice

- Use the template to create real data: set a strong precedent
- Grant funded project (*ca.* 2014):
  - create hybrid curation system: NLP & ML + expert curation
  - accelerate (rather than replace) scientists
  - gets easier and quicker as more assays are annotated
- Curated ~4000 assays from PubChem MLPCN (Molecular Libraries)
- Data can be found at:
  - http://beta.bioassayexpress.com
  - https://github.com/cdd/bioassay-express

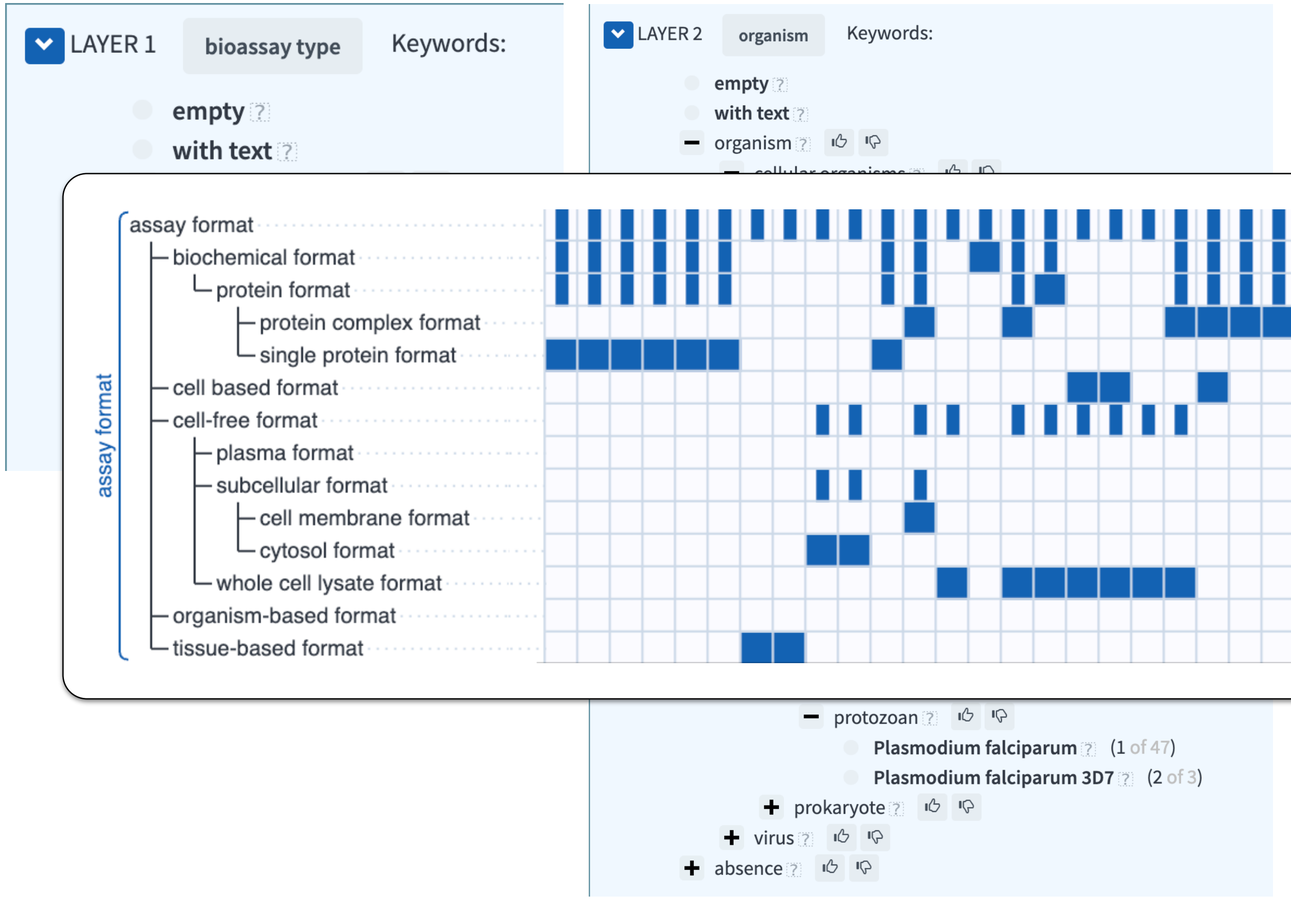
BioAssay Express

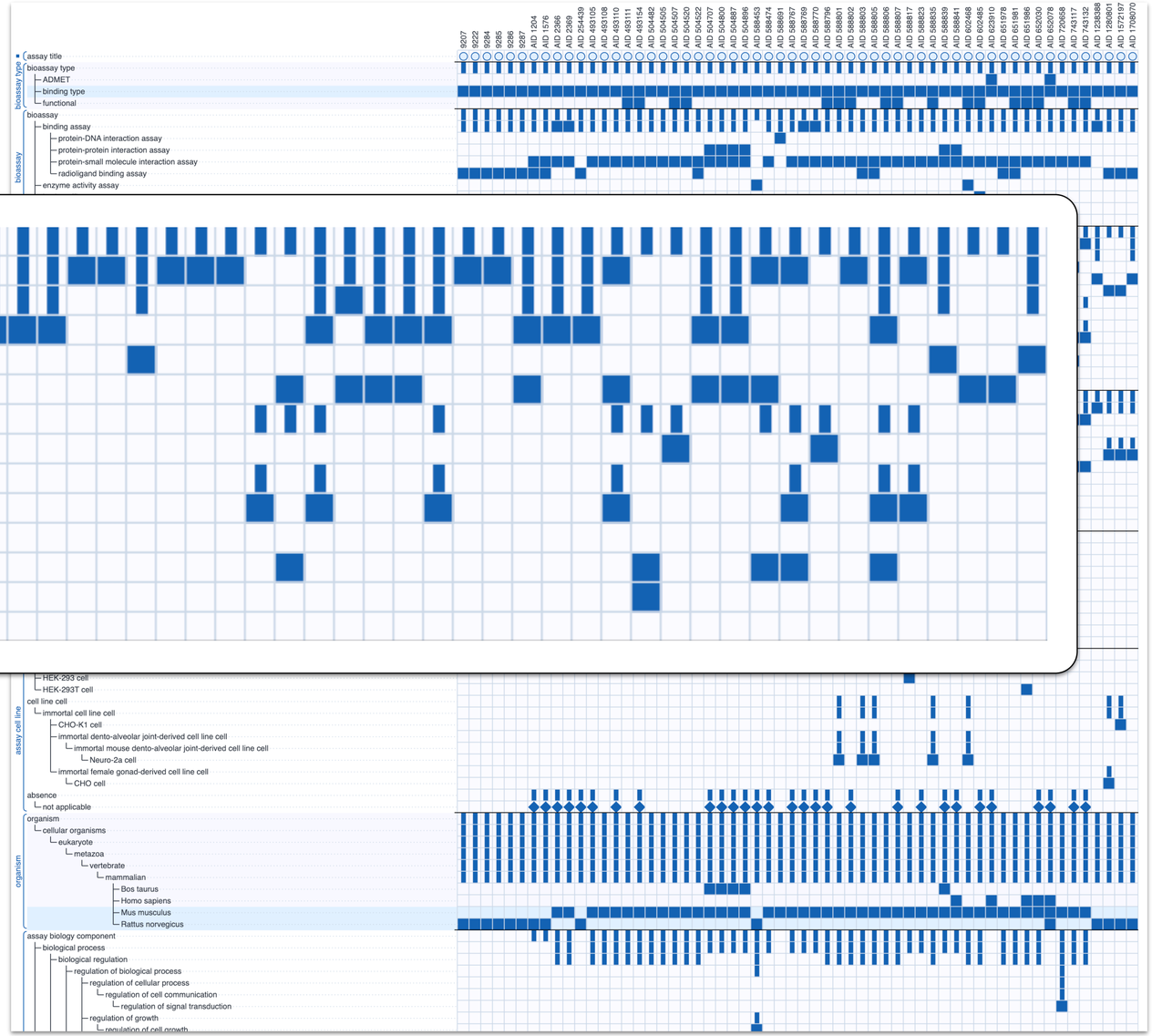
- Built for original curation of public data
- Test-bed for many experimental features
  - data entry
  - searching
  - analysis
  - machine learning
  - text creation
  - (among others)



What can you do with marked up assays?

- Search by narrowing with *exact* and *precise* criteria

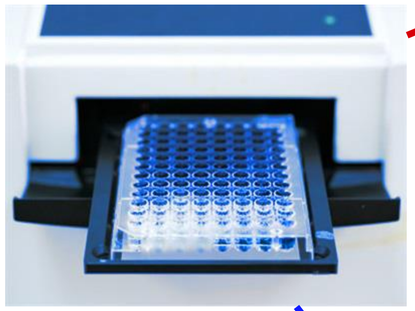


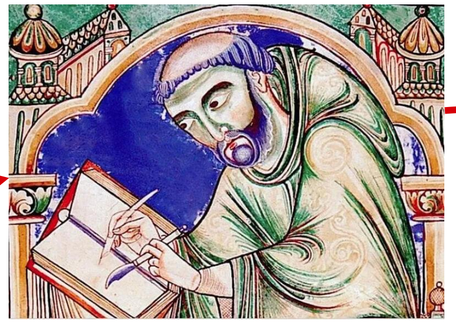


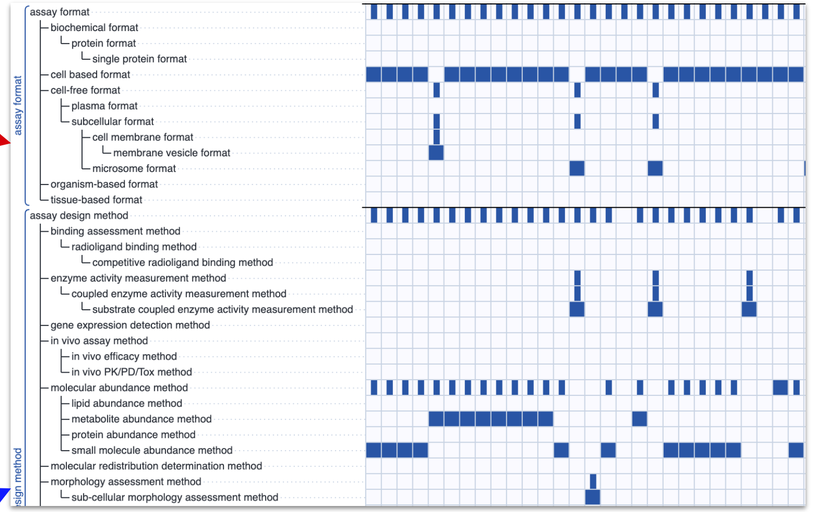
- Can also use annotations like fingerprints

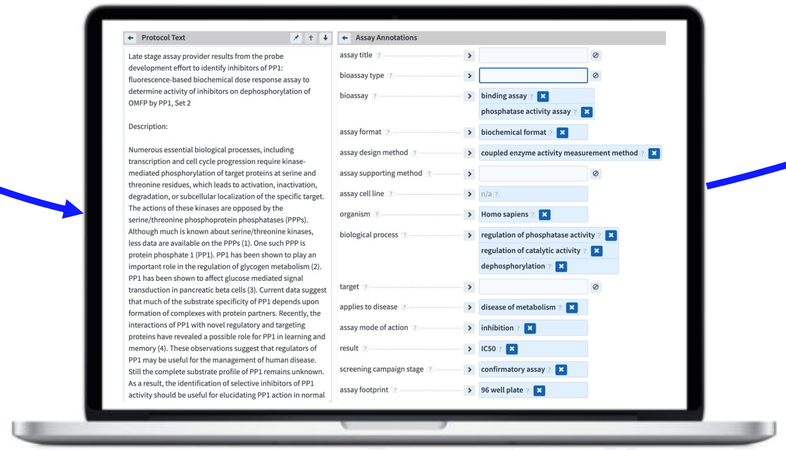
Curation: now vs. then

- Writing up text, then re-deriving machine readable description... ?





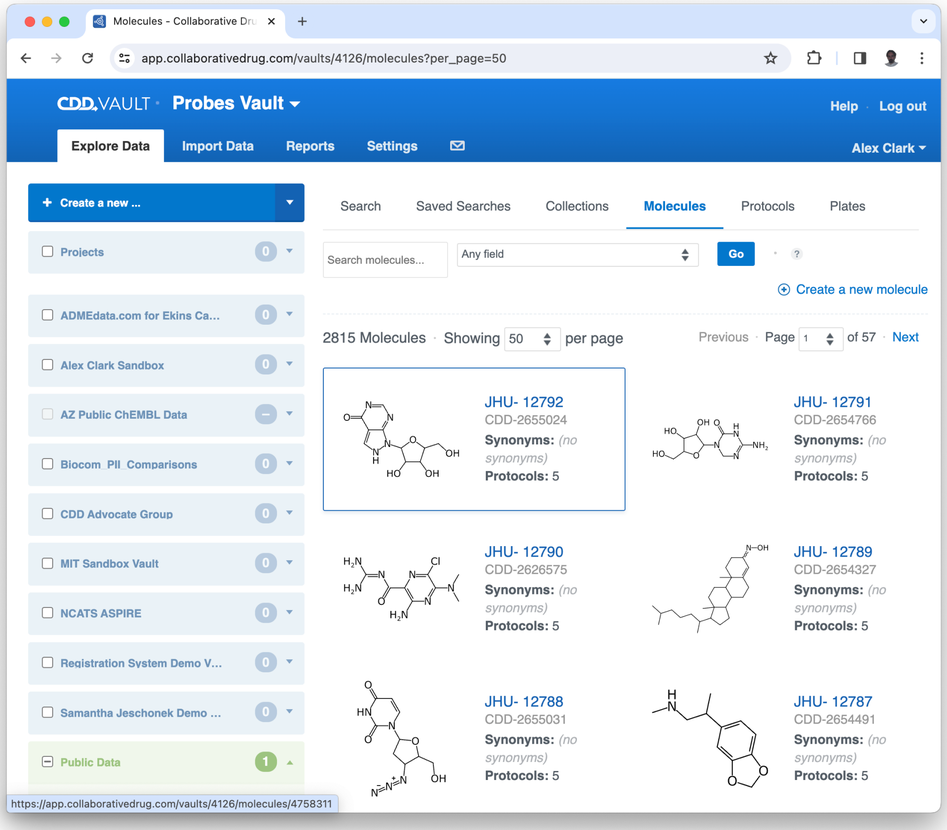




- Make the FAIR UI *easier*
- Introduce short term payoffs

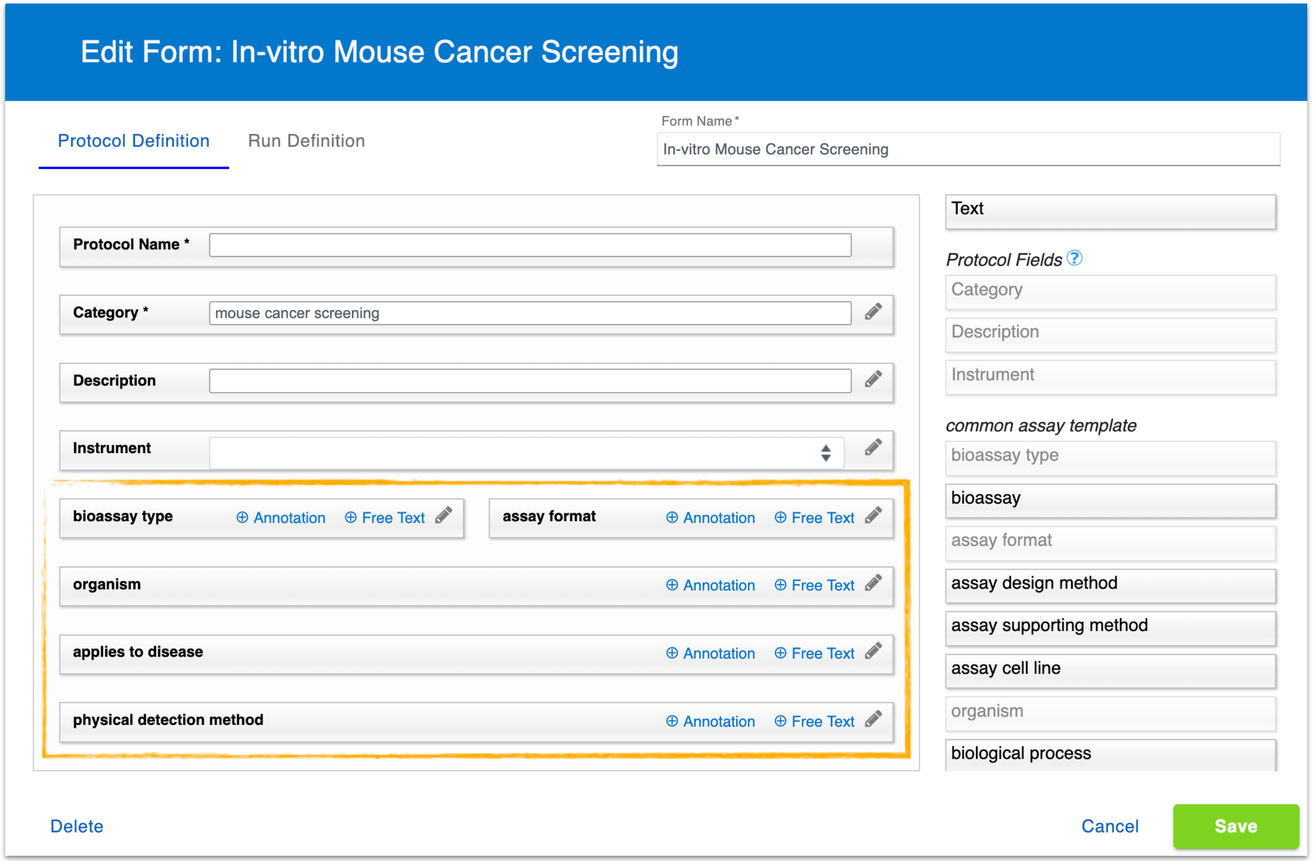
Annotation as a product

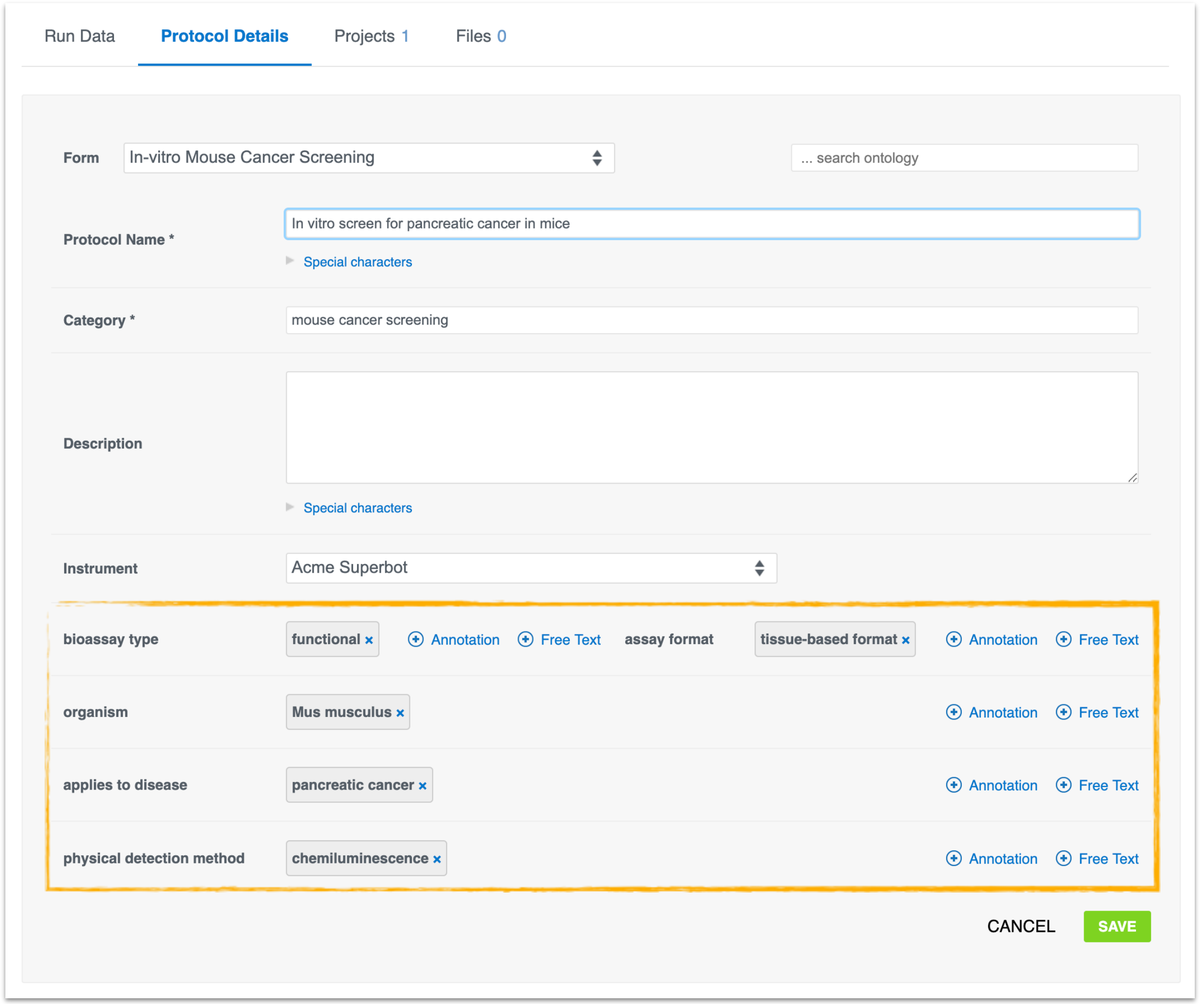
- **BioAssay Express**: advanced prototype, open source
- **CDD Vault Annotations**: 80/20 rule, same data model
- Vault is used daily by thousands of scientists, meet them where they're at
- Knowing about the *semantic web* is optional
- Introduce ontologies and templates gently
  - well defined set of terms
  - everyone has access to same vocabulary
  - like what they had before, except better



Protocol forms

- Mix text, numbers & pick lists with selected ontology terms



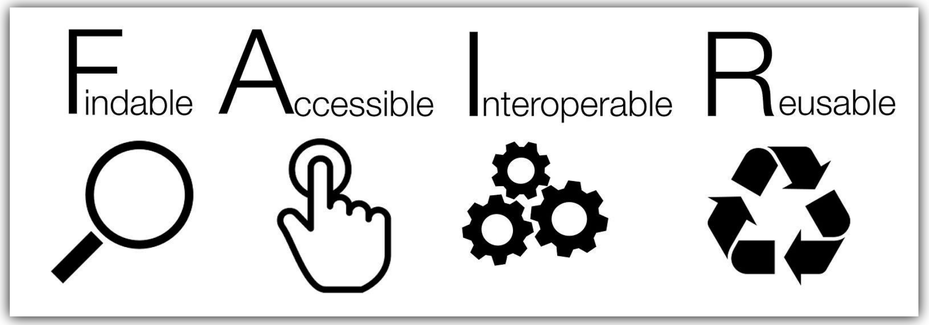


Business logic

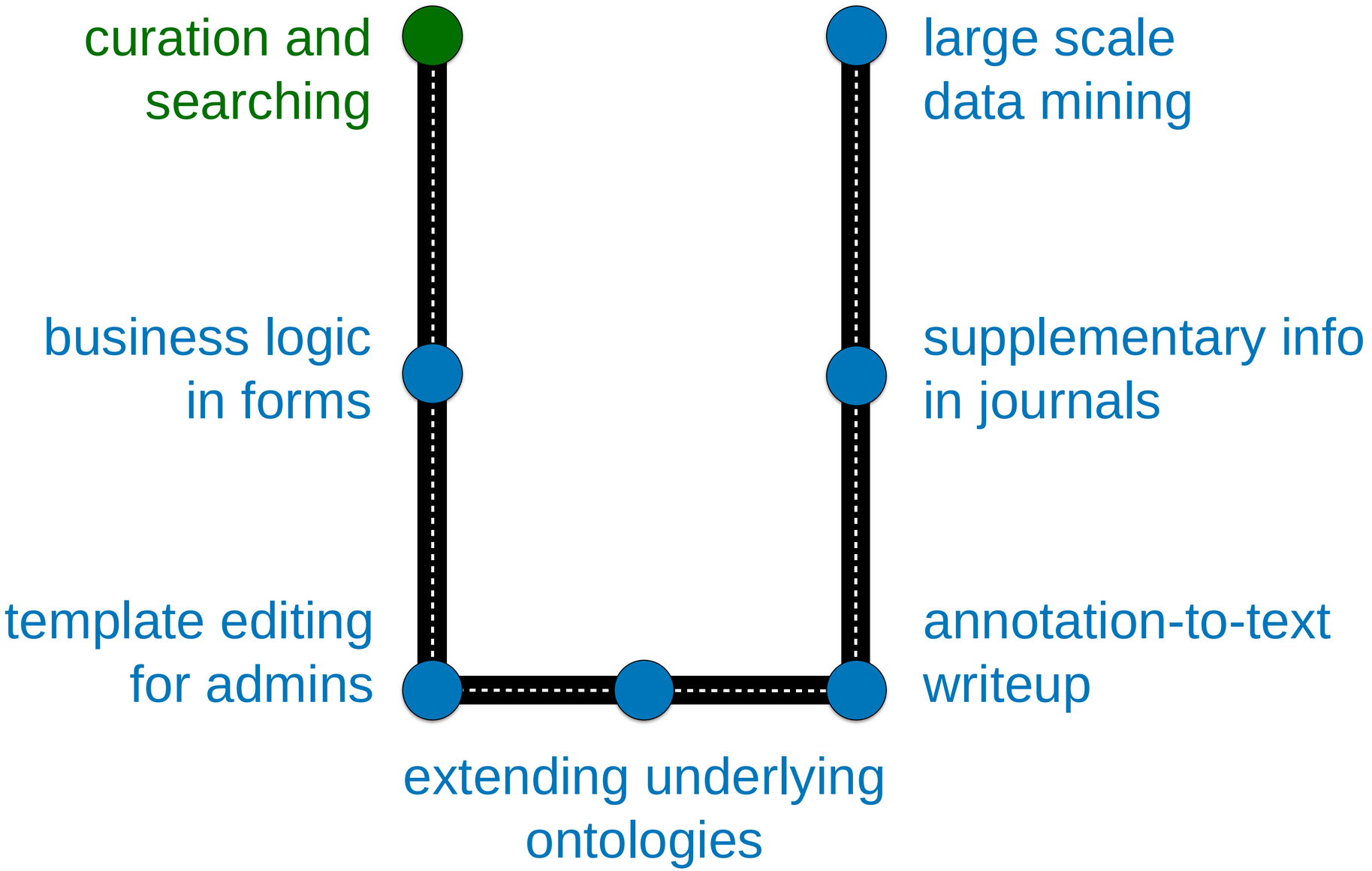
- Reduce curation efforts
- *and* fewer mistakes

Collaboration

- Common terms: ontologies are universal
- Assays are directly comparable, regardless of source
- Interoperability:
  - reduce/eliminate ambiguity
  - share assays with CROs (and acquisitions)
  - preserve institutional knowledge
- Completeness: you know what you know...



Roadmap


curation and searching
large scale data mining
business logic in forms
supplementary info in journals
template editing for admins
annotation-to-text writeup
extending underlying ontologies

Questions?

- Contact:
  - Alex M. Clark **alex@collaborativedrug.com** (Collaborative Drug Discovery)
- Thanks to the Vault & BioAssay Express teams