Corpora as evolving entities:
embedding corpora in biomedical ontologies

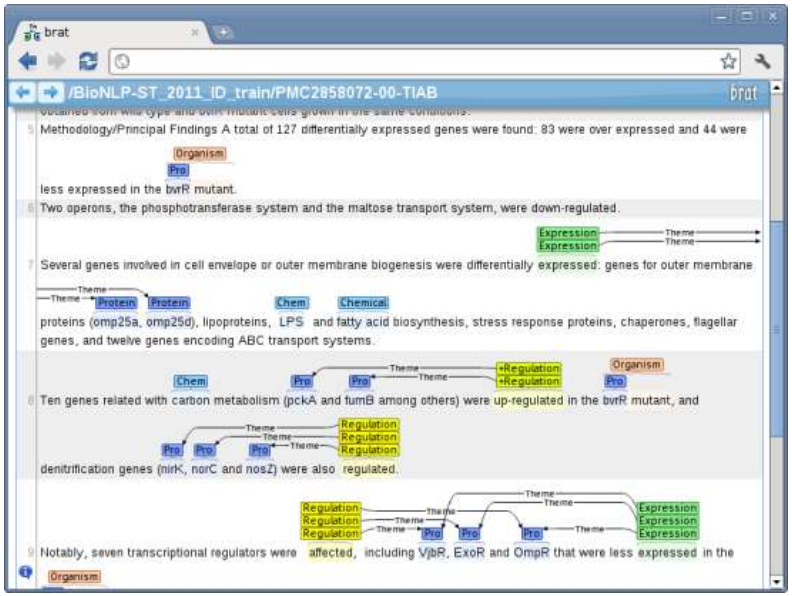
Elizabeth T. Hobbs[a], Stephen M. Goralski[a], Ashley Mitchell[a], Andrew Simpson[a], Dorjan Leka[a], Emmanuel Kotey[a], Matt Sekira[a], James B. Munro[b], Suvarna Nadendla[b], Rebecca Jackson[b], Aitor González-Agirre[c], Martin Krallinger[c,D], Michelle Giglio[b] & **Ivan Erill**[a]

[a] Department of Biological Sciences, University of Maryland Baltimore County
[b] Institute for Genome Sciences, University of Maryland School of Medicine
[c] Barcelona Supercomputing Center (BSC)
[d] Centro Nacional de Investigaciones Oncológicas (CNIO)

Biomedical corpora

- Essential resources
  - Map biomedical entities & concepts to scientific text
  - Enable training and benchmarking of text-mining systems

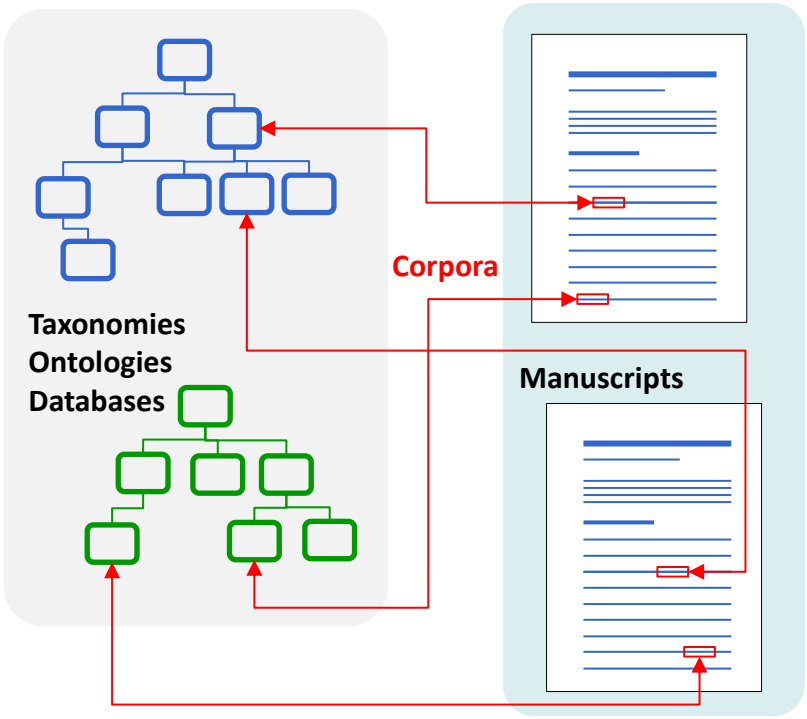
- FAIR issues
  - Specialization
  - Accessibility
  - Maintenance
  - Obsolescence



brat rapid annotation tool (https://brat.nlplab.org/)

Re-conceptualizing corpora

- Corpora as connectors
  - Provide mapping between object identifiers and manuscripts
  - Relationship can be stored on either end
    - Manuscripts are immutable
    - References (e.g. ontologies) evolve
  - References are logical candidates for corpora storage



The ECO-CollecTF corpus

- Annotation of evidence
  - Thematically focused
    - 87 open-access articles from CollecTF (bacterial TFBS database)
  - Evidence statements
    - a technique used in the context of making a claim about some entity
  - Entity categories
    - Gene Ontology (biological process, molecular function, cellular component)
    - Sequence Ontology
    - Ontology of Microbial Phenotypes
    - Taxonomy/phylogeny
  - Target text
    - Results and Discussion
    - Only single or consecutive sentences

The ECO-CollecTF corpus

Example annotations

a

PMID 24694298

"*S. lividans* AdpA directly regulates at least the six AdpA-dependent genes listed above and identified by microarrays and qRT-PCR analysis."

ECO:0000058

ECO:0001566

Biological Process

Resulting annotations

```
ECO:0000058 + Confidence: High + Biological Process + Assertion strength: High
ECO:0001566 + Confidence: High + Biological Process + Assertion strength: High
```

b

PMID 24694298

"These EMSA experiments demonstrated that *S. lividans* AdpA directly binds to five intergenic regions confirmed the in silico prediction presented in Table 2."

ECO:0000096

Molecular Function

Resulting annotation

```
ECO:0000096 + Confidence: High + Molecular Function + Assertion strength: High
```

c

PMID 4490131

"The endogenous protein expression and localization for WalR and WalK was also checked using confocal immunofluorescence microscopy. WalR could be localized to the cytoplasm of *B. anthracis*, but WalK could not be detected once again (data not shown)."

ECO:0005587

Cellular Component

Resulting annotation

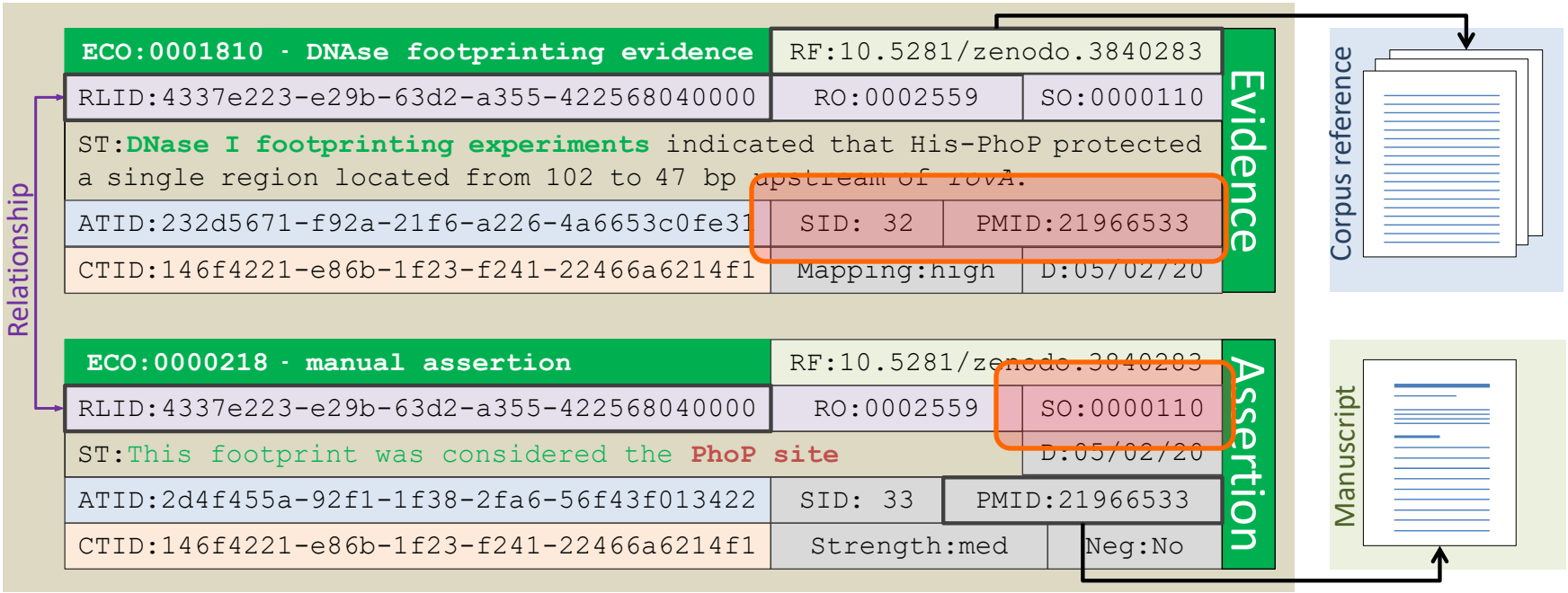
```
ECO:0005587 + Confidence: High + Sequence Feature + Assertion strength: High
```

Additional annotation fields

- Mapping confidence
- Assertion strength
- Negative assertion

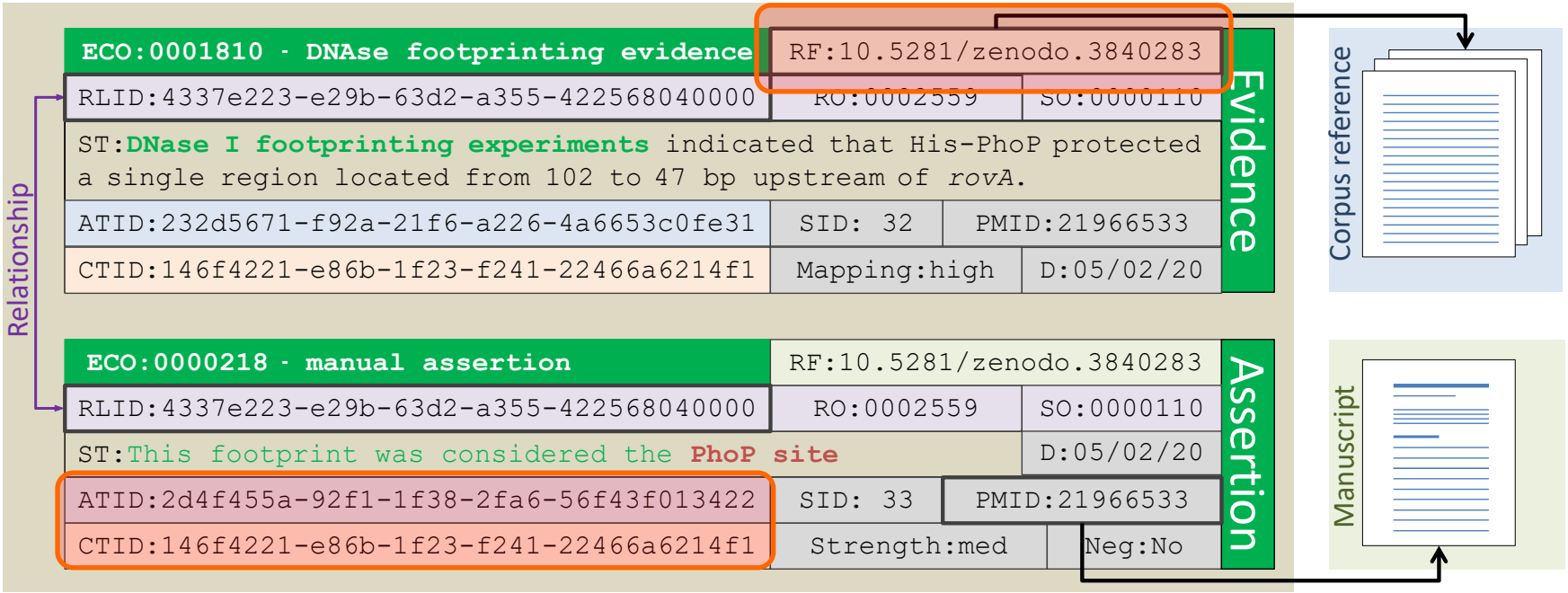
Ontology embedding

- Annotation property in ECO terms
  - Independent annotations for ECO *evidence* and *assertion* terms
  - Annotation identifiers
    - Sentence (SID) and PubMed (PMID)
    - Entity identifier (GO, SO, OMP)



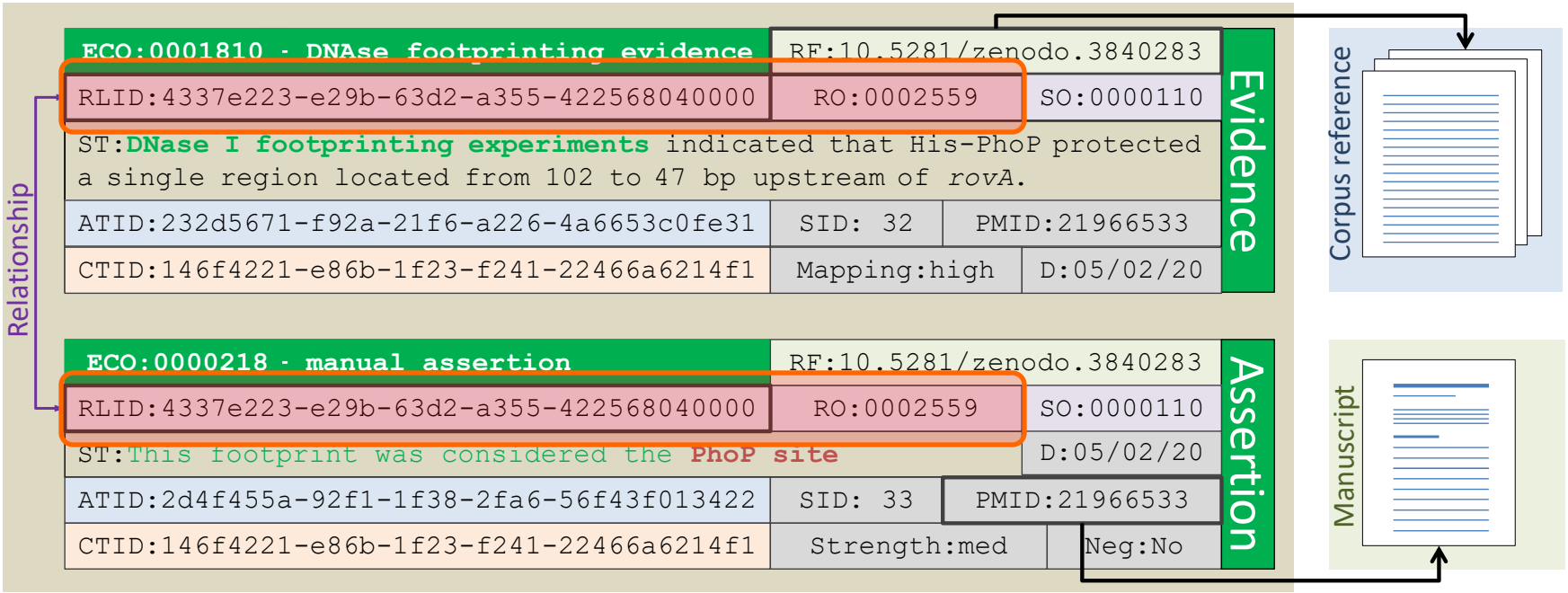
Ontology embedding

- Annotation property in ECO terms
  - Annotation identifiers
    - Corpus identifier (DOI)
      - Annotation guidelines, training material, parsing code
      - Pointers to dynamic curator table
    - Annotation and curator UUID identifiers (ATID and CTID)



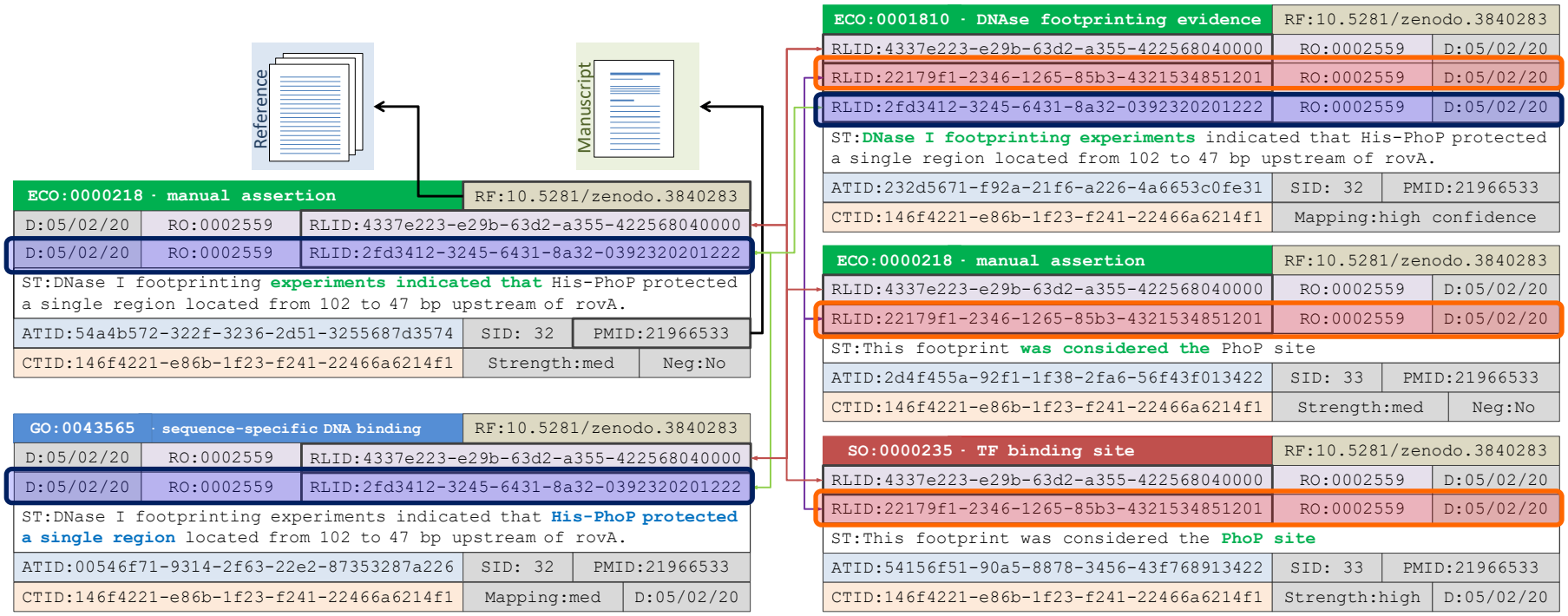
Ontology embedding

- Annotation property in ECO terms
  - Relationship array
    - Relationship type (RO)
    - Relationship identifier (RLID)
      - Binds multiple annotations



Ontology embedding

- Straightforward generalization to multiple ontologies
  - Multi-ontology
    - Each ontology embeds its annotations
  - Relationship array
    - Multiple relationships structure interrelationship across ontologies



Ontology embedding

- Advantages
  - Obsolescence
    - Corpus evolves with ontology
    - Backed up by long-term ontology funding
  - Findability & accessibility
    - Corpus accessible through ontology
    - Leverages ontology resources and community involvement
    - Ontology with examples of use that can be used contextualize terms
  - Reusability & interoperability
    - Annotations embedded independently
    - Standardized format
    - Implicit (term-based) links across corpora

Ontology embedding

- Future directions
  - OWL-based embedding of annotations
  - Development of a parsed sentence repository
    - Facilitate corpora embedding
    - Prevent congruence errors
    - Standardize parsing
  - Demonstrating feasibility of multi-ontology corpora
    - Ontology & corpus developer partners needed
    - Interested in participating?
      - Please contact
        - Ivan Erill (erill@umbc.edu)
        - Michelle Giglio (mgiglio@som.umaryland.edu)

Corpora as evolving entities:
embedding corpora in biomedical ontologies

Elizabeth T. Hobbs[a], Stephen M. Goralski[a], Ashley Mitchell[a], Andrew Simpson[a], Dorjan Leka[a], Emmanuel Kotey[a], Matt Sekira[a], James B. Munro[b], Suvarna Nadendla[b], Rebecca Jackson[b], Aitor González-Agirre[c], Martin Krallinger[c,D], Michelle Giglio[b] & **Ivan Erill**[a]

[a] Department of Biological Sciences, University of Maryland Baltimore County
[b] Institute for Genome Sciences, University of Maryland School of Medicine
[c] Barcelona Supercomputing Center (BSC)
[d] Centro Nacional de Investigaciones Oncológicas (CNIO)

In memoriam
Jim Hu
Scientist, mentor, friend