SciBite

an ELSEVIER company
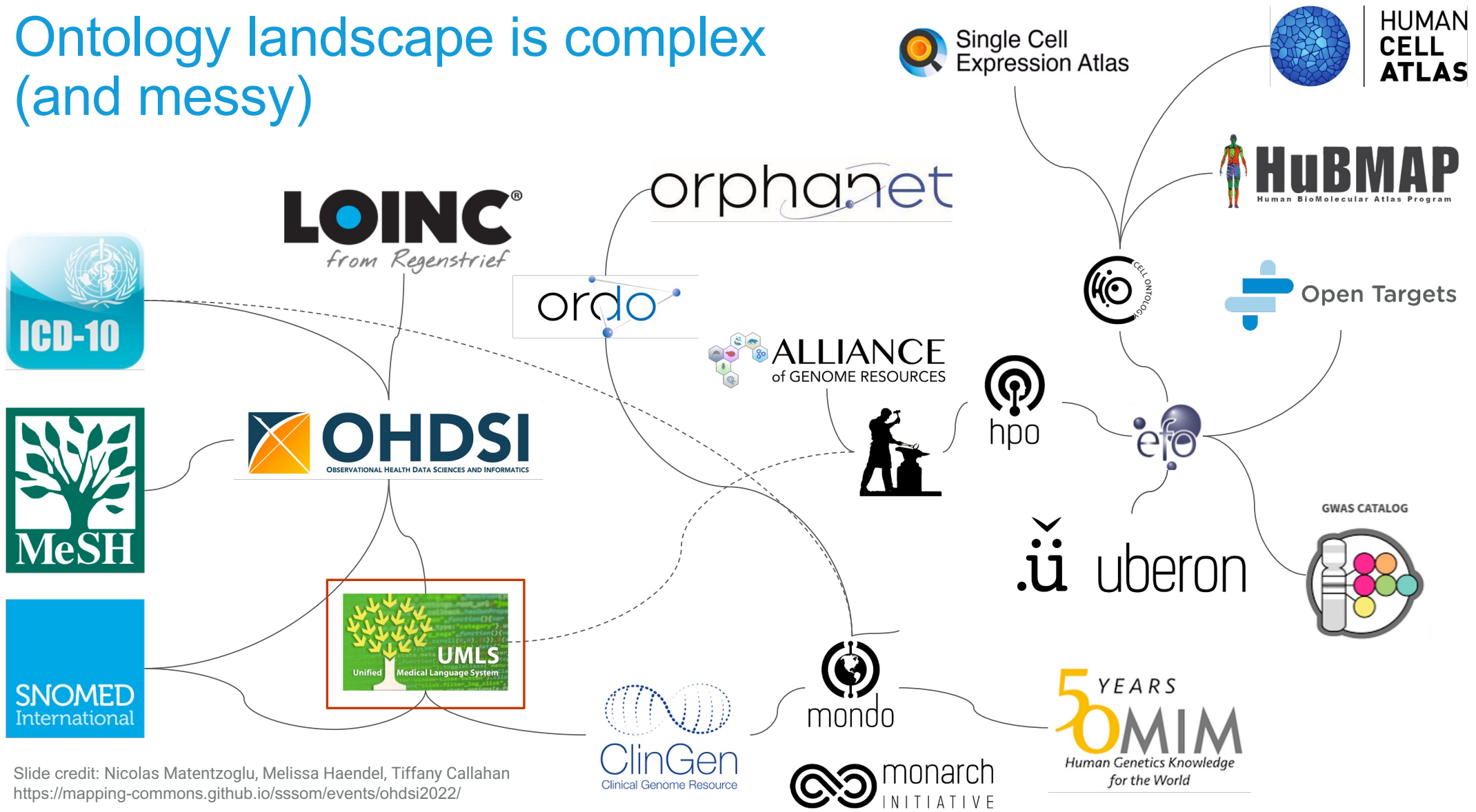
Managing mappings at scale

Simon Jupp

Head of Semantic Technology

Ontology landscape is complex (and messy)

Mappings from public sources

- More than 10 million extracted

- Often have weak semantics – many use *has_dbxref*

- Prefixes of CURIEs ambiguous – need normalization (HT Bioregistry and https://identifiers.org/)

- Hard to judge quality

Available ontology mappings

- Many mappings available in the public domain provided by terminologies e.g.

  - MedDRA->SNOMED CT
    (https://www.meddra.org/software-packages)

  - MedDRA->ICD10
    (https://www.meddra.org/software-packages)

  - SNOMED CT->ICD10
    (https://www.snomed.org/maps)

  - SNOMED CT to ICD-O
    (https://www.snomed.org/maps)

Mapping challenges

Locating mappings

Normalizing mappings

Provenance/quality

Custom mappings

Governance
Versioning
Updates

Overview of Semi-Automated Ontology matching approaches

Rule-based

- Lexical mapping e.g., Agreement Maker Light (AML)
  - Looking at label and synonym string similarity
- Structural/logic based e.g., LogMap
  - Exploit the ontology structure and relationships to infer mappings

Machine learning

- Word embeddings e.g., word2vec
- Language models e.g., BERTMap, MapperGPT

Ontology matching for life sciences

- Generally combing approaches work best
- Existing tools can generate large amounts of "likely" mappings

Challenges

- Difficult setup and installation of tools
- Poor performance over large terminologies
- Lack of emphasis on curation, versioning and maintenance
  - Lack a good user interface
- Subtly and complexity of life-science vocabulary e.g. indication vs phenotype vs adverse event
  - Semantics of what constitutes a match is often context dependent

SciBite mapping service

- Normalize mappings derived from ontologies

  - Add reciprocal mappings

  - Provide SSSOM-compliant metadata so they can be filtered

- CURIE-index based on Bioregistry

- Supplemented with value public sets

- SciBite-curated sets to fill gaps

- Access via SciBite tools – CENtree, SciBiteSearch

- Supplement with custom mappings from WorkBench

SciBite Workbench

- Platform for generating mappings at scale

- Utilizing SciBite's NER technology, public sources and SciBite curated vocabularies to predict and score mappings

- Allows curators to review and validate mappings prior to publishing

- Easy to use and manage versions and provenance of mappings

Utilising SciBite's Mediating Vocabularies

- Utilising SciBite extensive curated vocabularies and public sources to act as mediators when generating mappings between source and target ontologies that have very little metadata (synonyms on terms)

  - You can supply a mediating vocab to help find more mappings

  - E.g., When mapping two anatomy ontologies e.g., MA and NCIt anatomy, we can use SciBite Anatomy vocab to help discover potential matches

SciBite Anatomy vocab (SB:12345)

"Auditory Ossicle"

"auditory bone"

MA term (MA:0000254)

"auditory bone"

"Ear Bone"

NCI Term (NCI:C32164)

"Auditory Ossicle"

"Bone of the Tympanum"

No direct match found but both match to same term in mediating vocab
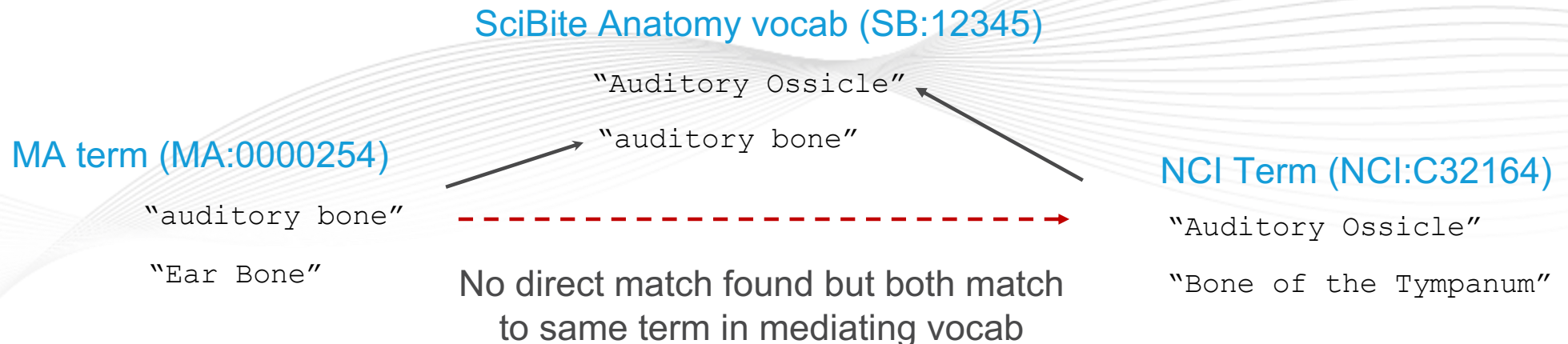
Creating an ontology alignment in SciBite Workbench

Workbench *Curate and annotate*

Add New Mapping

Mapping Name
Mouse (MA) to Human (NCIt) Anatomy
34 / 100

Source Ontology
MOUSE_MAPPING

Target Ontology
HUMAN_MAPPING

Mask Terms

vein (180) ⊗ | artery (166) ⊗ | muscle (149) ⊗ | nerve (107) ⊗ | nucleus (106) ⊗ | ✕ SUGGEST MASK TERMS

intestine (36) ⊗ | spinal (36) ⊗ | ganglion (36) ⊗ | hand (36) ⊗ | lower (37) ⊗

foot (37) ⊗ | renal (41) ⊗ | cortex (41) ⊗ | vertebra (44) ⊗ | left (44) ⊗ | skin (45) ⊗

dorsal (46) ⊗ | respiratory (47) ⊗ | digit (48) ⊗ | connective (48) ⊗ | right (49) ⊗

smooth (51) ⊗ | layer (52) ⊗ | duct (54) ⊗ | blood (54) ⊗ | vessel (59) ⊗

tissue (66) ⊗ | epithelium (85) ⊗ | system (87) ⊗ | bone (94) ⊗ | gland (97) ⊗

☑ Best Hits only

☐ Exact Match only

Cutoff score
0.7

Mediating Ontology
Anatomy   ✕ SUGGEST MEDIATING ONTOLOGY

CANCEL   GENERATE MAPPING

Workbench identifies potential masking terms

Workbench will identify SciBite vocabularies that will act as mediators

Can quickly filter on Exact, Broader or No match

Explore mappings of different quality

Publish your mappings to users

Publish your mappings to our Ontology Management System (CENtree) with provenance