eurofins | DISCOVERY

# Describing the Need: An Ontology End-user Case Study

Sheryl Denker & Antal Berenyi

Workshop on Synergizing Biomedical Ontologies

July 16, 2021

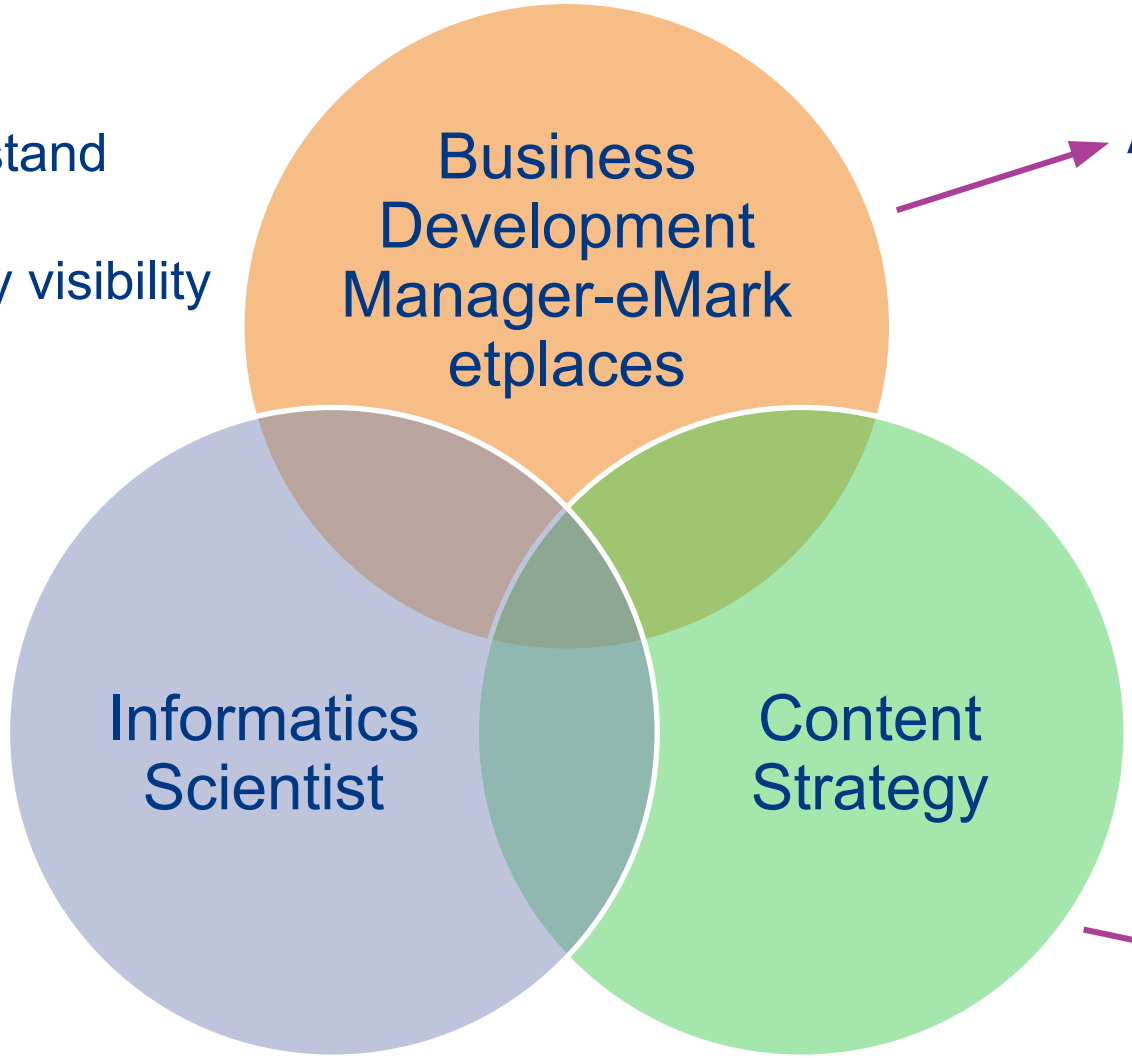
A STRONG FOUNDATION FOR SUCCESSFUL DRUG DISCOVERY

PHENOTYPIC CENTER OF EXCELLENCE

# Ontology Consumer: Life Sciences Products & Services CRO

eurofins | DISCOVERY

Sales Goals:
- customers understand what we offer
- increase company visibility and sales




Aimee

Data Science Goals:
- increase and improve data usability
- team efficiencies

Antal

Content Marketing Goals:
- harmonize descriptions
- findable across platforms

Sheryl

# The Challenge of eMarketplaces

eurofins | DISCOVERY

- ~1500 provided terms manually curated by 6 employee contributors (Product/Content Managers)
- Tedious effort resulted in a list of 523 terms
- A second, independent pool of 5K terms was provided
- How can we:
  - ✔ reduce manual labor in the curation effort
  - ✔ obtain only highly relevant results
  - ✔ not miss important terms

| | | | | | |
|---|---|---|---|---|---|
| Leukemia Models | Capillary Electrophoresi: | Human Immortalised Cell C | Ocular Permeability | Biosensor Assays | Histone Deacetylase (HDAC) |
| Thrombosis Models | Cell Free DNA Assay | Human iPSC's | Skin Permeability | Bone Resorption Assay | In Vitro Enzyme Activity Assay |
| Zucker Rat | Digital Polymerase Chai | Human Adult Stem Cells | Vaginal Permeability | Catecholamine Assay | Kinase Assays |
| Bone Metastases Models | DNA Pull Down Assay | Human MSC's | Cell Fractionation | Cell Based Assays Consultir | Methyl CpG Binding Domain ( |
| Bone Models | DNA-lipid interaction an | Human HSC's | Compound Distribution | Cell Membrane Lipidomics | Methyltransferase Assays |
| Cartilage Repair Models | DNase Footprinting Ass | Human DNA | Drug Distribution Studies | Chorioallantoic Membrane ( | Mutase Assays |
| In Vitro Bone Models | Electrophoretic Mobility | Human RNA | PharmacoScan - A Pharm | Chromium Release Assay | Nitric Oxide Synthase Assays |
| Osteoarthritis Models | Isothermal Amplificatior | Human Cell Lysates | Red Blood Cell (RBC) Pa | Ciliary Motility | Oxidase Assays |
| Osteoporosis Models | Linear DNA Amplificatio | Human Cell Products | Conjugate Detection | Co-cultivation Microscopy A | Oxygenase Assays |
| Artery On-A-Chip Platform | Plasmid Retention Assa | Human Whole Blood | Glucuronyltransferase | Co-Culture Angiogenesis As | PARP Assays |
| Atrial Arrhythmias Models | Polyacrylamide Gel Elec | Human Umbilical Cord Bloo | Glutathione Transferase | Coagulation | PDZ Domain Assays |
| Cardiovascular Metabolics Mod | qPCR Analysis | Human Serum | N Acetyl Transferase | Cytokine Release Assay | Peroxidase Assays |
| Cardiovascular Models | Recombinant Nucleoso | Human Plasma | Sulfotransferase | Cytokines | Phosphatase Assays |
| Coronary Artery Disease Models | Reverse Transfection | Human RBC's | CYP Induction | Ecarin Clotting Time (ECT) | Phosphodiesterase Assays |
| ExeGen® LDLR MiniSwine | Threshold DNA Detectic | Human Buffy Coat | CYP Inhibition Assay | Factor X Assay | Phosphorylase Assays |
| In Vitro Cardiovascular Models | Tissue Microarray | Human PBMC's | CYP Isozyme Mapping | Fibrinogen Assays | Polymerase Assays |
| In Vitro Hypoxia Models | Whole Genome Amplific | Human Leukopacks | Cytochrome P450s | Glucose Release Assay | Protease Assay |
| Ischemia Models | Amplicon - HLA Typing | Human Biofluids | Corning® HepatoCells | Glucose Uptake | Racemase Assays |
| Japanese White Rabbit | Amplicon 16s rRNA Seq | Human synovial fluid | Drug Drug Interactions | Hairless Mouse | RNA Methylation Assays |
| Myocardial Infarction Models | ChIP-sequencing Profili | Human semen | Phase I Metabolism | High Content Screening | Synthetase Assays |
| Ossabaw-ptFH3 Patient Model™ | De Novo Genome Sequ | Human vaginal fluid/swap | Phase II Metabolism | High Throughput Screening | Ubiquitin Ligase Assays |
| Pulmonary Hypertension Animal | Exome Sequencing | Human amniotic fluid | GI Fluid Stability | Histamine Release Assay | UDP-glucuronosyltransferase |
| Restenosis Models | Metagenomic Sequencir | Human ascites | Metabolic Stability | In Vivo Angiogenesis Assay | Anaplastic lymphoma kinase ( |
| SHR/NCrl Rat | mtDNA Sequencing | Human Vitrious Humour | Metabolic Stability in Hep- | Inflammation Assays | Angiopoietin-2 (Ang-2) Assays |
| Ventricular Tachycardia Models | NGS: Next Generation S | Human Tears | Metabolic Stability in Micr | Interferon Gamma Release A | B-Cell Lymphoma 2 (BCL-2) A |
| White Rabbit | RAD Sequencing | Human Sputum | Plasma Stability Assay | International Normalized Rati | Carcinoembryonic Antigen (C |
| Zucardio | Sanger Sequencing | Human gastrointestinal fluid | Isotopic Labeling | Lipid Detection | CD134 (OX 40) Assays |
| Zucker Rat | Targeted Sequencing/Re | Human Bone Marrow aspira | Metabolite Generation | Lipid Disorder Screening | CD137 Assays |
| Acute Kidney Failure Models | TCR Repertoire Sequen | Human Faecal Matter | Metabolite Identification | Lipid Kinase Assays | CD20 Assays |
| Acute Liver Failure Models | Whole Genome Bisulphi | Human Nasal Fluid | Metabolite Quantification | Lipid Peroxidation Assay | CD252 (OX 40L) Assays |
| Acute Lung Injury Models | Whole Genome Sequenc | Human BronchoAlveolar La | Metabolite Structural Eluc | Lipolysis | CD79b Assays |
| Critical Care Models | ChIP-sequencing Profili | Human Breast Milk | Metabolite Synthesis | Low Density Lipoprotein (LDL | Checkpoint Kinase-1 (ChK-1) A |
| Emesis Models | Expression Profiling | Human Bile | NMR Methods | Matrigel Plug Assay | Colony Stimulating Factor 1 R |
| Liver Disease Animal Models | RNA Profiling | Human Saliva | No Label Methods | Metabolomics | Cytotoxic T-Lymphocyte Antig |
| Liver Fibrosis Models | Target Gene Expression | Human Urine | Non Isotopic Labeling | Micromanipulation Microsco | Extracellular signal Regulated |
| Sepsis Models | Whole Genome Profiling | Human cerebral spinal fluid ( | Radiolabeled Tracers | Oxidative Burst | Fibroblast Growth Factor Rec |
| Tissue Rejection Models | CRISPR | Human TMA | Reactive Metabolite | Pathway Mapping | Human Double Minute 2 (HDM |
| Zucker Rat | Engineered Meganuclea | Human custom TMA | Reference Standards | Phenotype | Human Epidermal growth facto |
| Acne Models | Genome Editing | Human Brain | Equilibrium Dialysis | Phenotypic Assays | Human Organoids |
| Atopic Dermatitis Models | Recombinant AAV Med | Human Tonsils | Protein Aggregation | Platelet Function Testing | Hypoxia-Inducible Factor 1 (HIF |
| Dermatology Models | Transcription Activator I | Human Eyes | Protein Binding HPLC | Proteomics | Immuno-oncology Assays |
| Hair Growth Models | Zinc Finger Nuclease Ge | Human Ear | Protein Binding Ultrafiltrat | Reactive Species Assays | Indoleamine 2,3 Di-Oxygenase |
| Hairless Mouse | Allele Specific Oligonucl | Human Hair | | Serotonin Release Assay | Kallikrein (KLK)-related Serine |
| Hairless Rat | Amplified Fragment Len | Human Larynx | | Sprouting Assay | Killer cell Immunoglobulin-like |
| In Vitro Skin Models | DNA Hydroxy Methylatic | Human Head and Neck Samples | | Stem Cell Differentation Mic | Liposome Immunoassay |
| Lupus Models | Genotyping | Human Lung | | Synthetic Lethality Screening | Lymphocyte Activation Gene- |
| Pruritus (Itch) | Random Amplified Polyı | Human Heart | | Thrombin Generation Tests | Lysine Specific Demethylase-1 |
| Psoriasis Tissue Models | Restriction Fragment Le | Human Vessels | | Thrombophilia Screening | Mammalian Target of Rapamy |
| Rosacea Models | Artificial Chromosomes | Human Breast | | Tube Formation Assay | Mesothelin Assays |

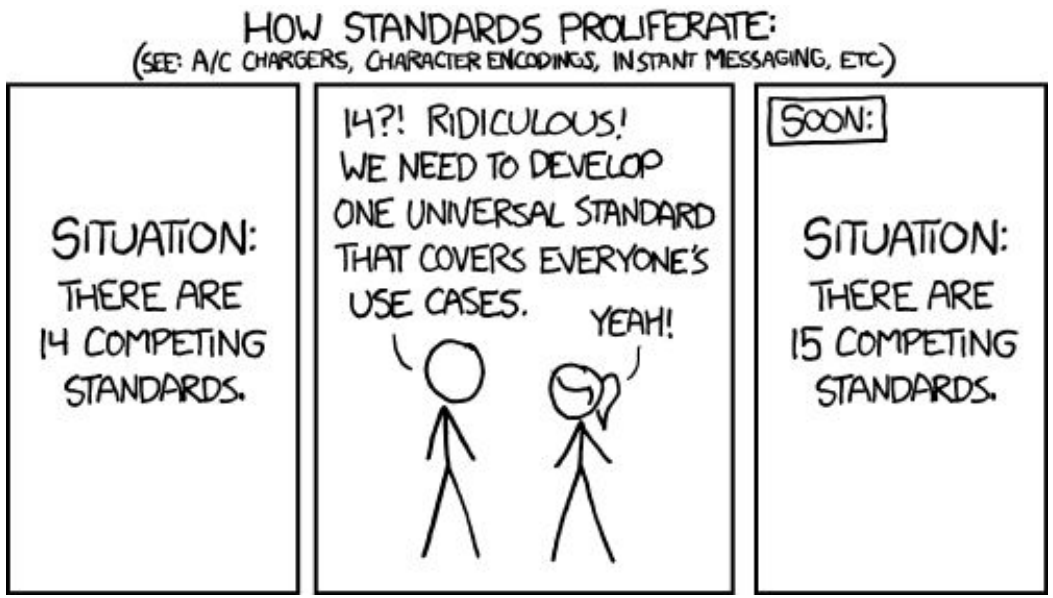
# Use An Ontology! But Which One?

eurofins | DISCOVERY

~900 ontologies related to biomedical sciences

Search repository https://bioportal.bioontology.org

- cell-based assays
- phenotypic assays

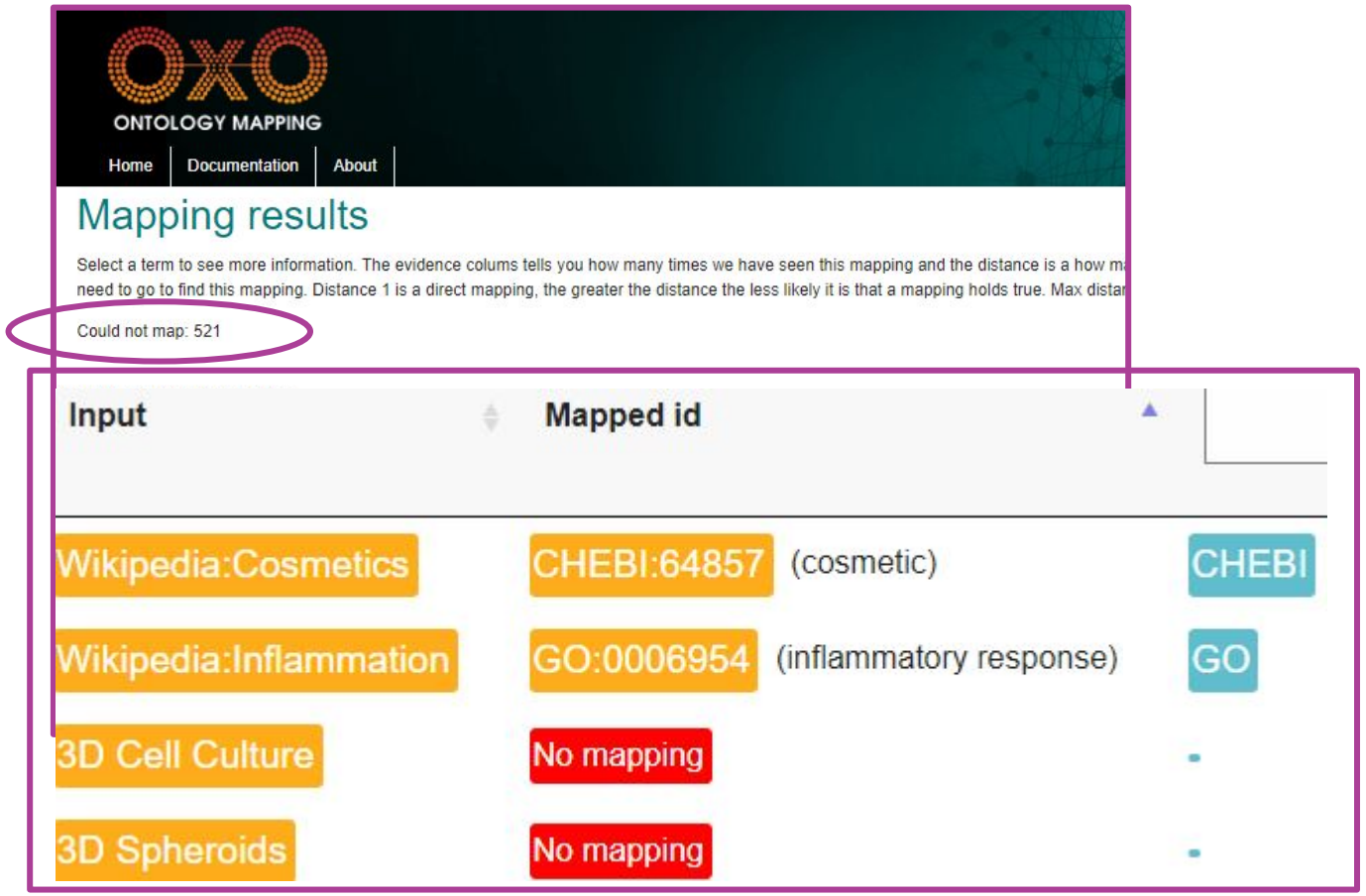



https://xkcd.com/927/

20-30 results, including BioAssay Ontology, SNOMEDCT, GO-PLUS

Where do we go from here?

# Let's Use A Mapping Tool!

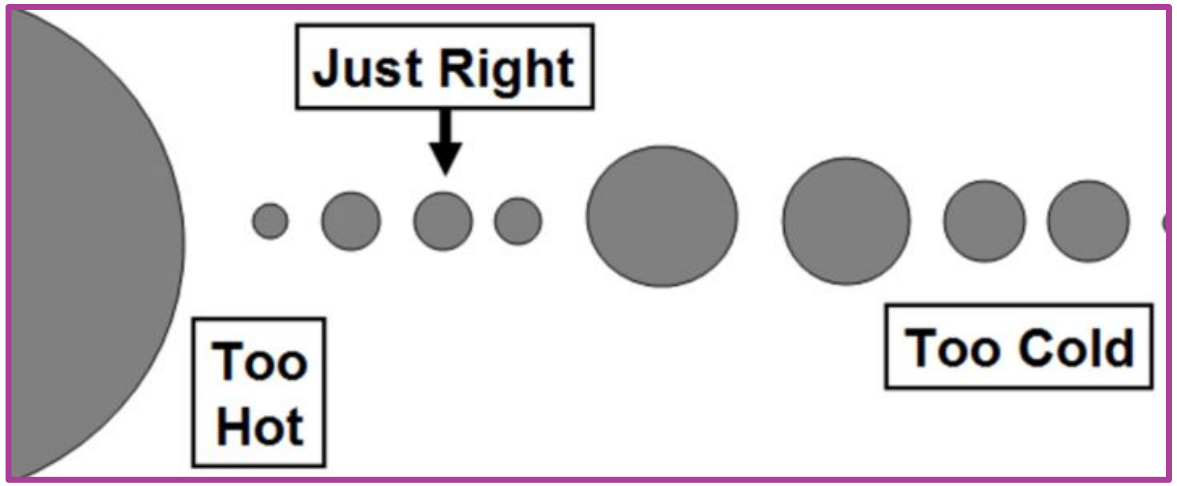



- OxO finds cross-references between ontologies, **vocabularies** and coding standards
- Only 2 out of 523 terms were found: inflammatory response and cosmetic
  - Chemical Entities of Biological Interest (CHEBI)
  - Gene Ontology (GO)

Finding and using the correct ontology is challenging for end users.
Our ultimate goal is to contribute assay data to an ontology that we find helpful.

# Establish Our Own Search Criteria

1. Biology and Users
   - ✔ terms & classes (cell-based assays)
   - ✔ fit for commercial purposes
2. Complex
   - ✔ Large number of assay terms to cover breadth of offerings from Eurofins Discovery
3. Maintenance
   - ✔ Currently active and maintained
   - ✔ Identified owners




Baum, Seth. (2013). *Journal of Sustainability Education.*

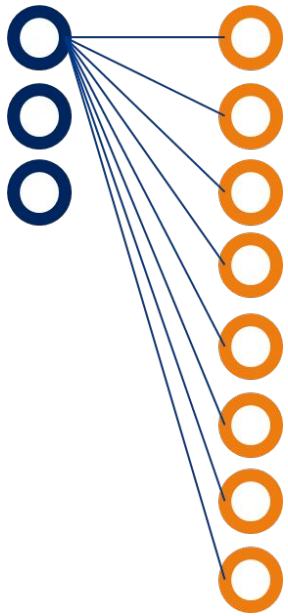
Interesting, but minimally helpful in getting our work done.

# Get A “Simple” Task Done: Accurately Match Complex Concepts

eurofins | DISCOVERY

Sample terms to match:

- *affinity tag purification*
- *affinity binding assays*
- *inflammation assays*
- *assays for inflammation*



- ~500 terms from one eMarketplace against ~5K terms from another
- ~2.5M pair-wise comparisons by eyeballing
- one pair/second…that’s about 30 days non-stop

Needs a more analytical approach.

# Enter Data Science

eurofins | DISCOVERY

Approaches in increasing usefulness

- Eyeball
- Excel VLOOKUP
  - Exact match
  - Approximate match
- NLP
  - Levenshtein distance
  - Cosine similarity

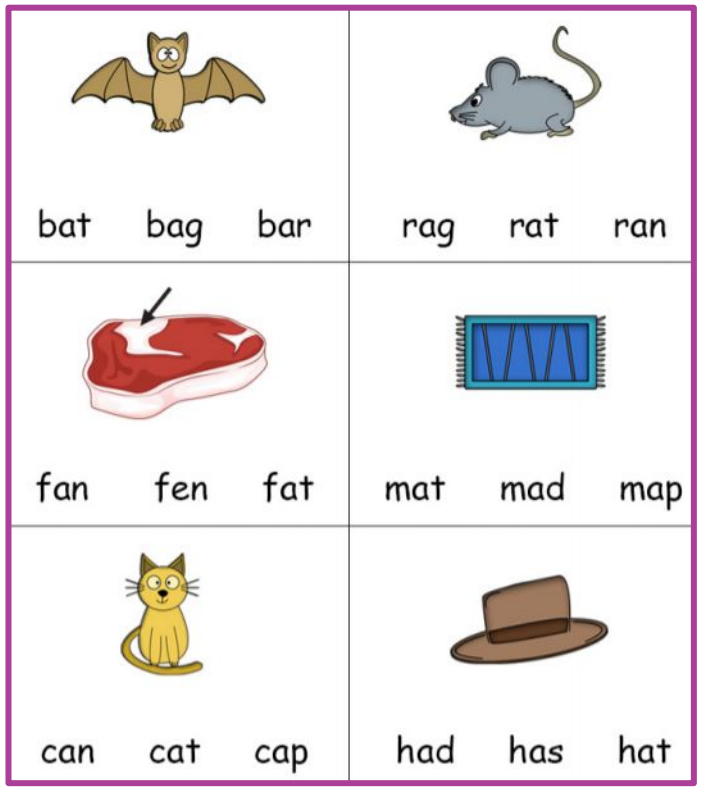
# A Better Approach than Eyeballs?

eurofins | DISCOVERY

## Excel VLOOKUP options

- Exact match - matches exact terms well
- Approximate match - only looks at matching characters at the beginning of the words
  - catapult and caterpillar will be flagged as matches!

In cases where there are few exact matches, Excel doesn't help much. What about approximate matches?




https://www.myteachingstation.com/

# Approximate Matches Produced by Excel VLOOKUP

eurofins | DISCOVERY

| Search Term | Approximate Match |
|---|---|
| Basic Metabolic Profiling | Basic Metabolic Panel |
| Behavioral Testing Models | Behavioral Phenotyping |
| Bioequivalence Studies | Biodistribution Studies |
| BioMAP® Human Phenotypic Platform | Biology |
| Biomarker Services | Biomarker Discovery |
| Biomarker Services Consulting | Biomarker Discovery |
| Biosafety Services | Bioproduct Safety Testing |
| Biotin Protein Conjugation | Biostatistics & Bioinformatics |
| Biotin-Antibody Conjugation | Biostatistics & Bioinformatics |
| Bone Models | Bone Mineral Density Testing |

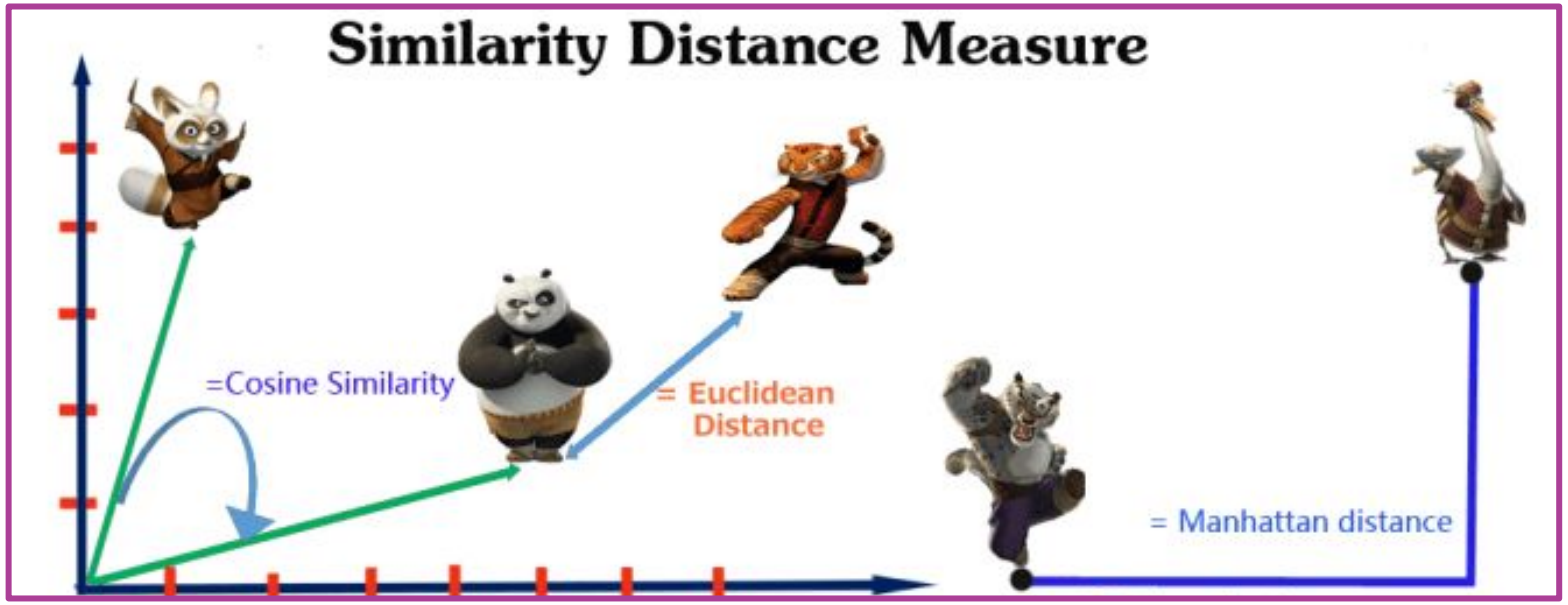
# A Better Approach: Computational String Comparison

eurofins | DISCOVERY

- Comparing words, phrases, terms – collectively called strings – is a non-trivial task
- Finding exact matches is easy, but finding a degree of similarity between strings is difficult
- A change in a single character can signal a small or a large difference in meaning:
  - Assay : assays – similar meaning
  - 3d cell culture : 2d cell culture – different meaning
  - Apparently different strings can carry the same meaning:
    human – homo sapiens

# Comparison Metrics

- Various metrics have been developed to measure similarity between terms
- We have used two similarity metrics and compared their usefulness
  - Levenshtein distance
  - Cosine similarity




https://dataaspirant.com/five-most-popular-similarity-measures-implementation-in-python/

# First Metric: Levenshtein Distance

- The minimum number of single-character edits (insertions, deletions or substitutions) required to change one word into the other

- For example, the Levenshtein distance between "kitten" and "sitting" is 3:

1. kitten → sitten (substitute "s" for "k")
2. sitten → sittin (substitute "i" for "e")
3. sittin → sitting (insert "g")

https://en.wikipedia.org/wiki/Levenshtein_distance

# Small Distances Don't Always Mean Closely Related

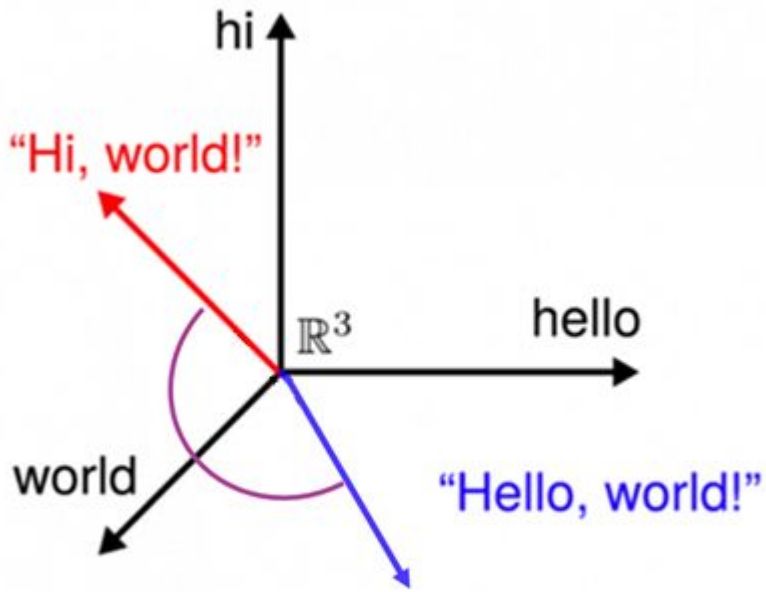
eurofins | DISCOVERY

| Query Term | Match_1 | LSim_1 |
|---|---|---|
| immunophenotyping | immunophenotyping | 0 |
| 3d cell culture | 2d cell culture | 1 |
| caco2 cell line | caco2 cell lines | 1 |
| caspase assay | caspase assays | 1 |
| cell based assays | cellbased assays | 1 |
| gene set enrichment analyses | gene set enrichment analysis | 1 |
| human buffy coat | human buffy coats | 1 |
| protease assay | protease assays | 1 |
| pulmonary hypertension animal model | pulmonary hypertension animal models | 1 |
| rna extraction | dna extraction | 1 |
| western blots | western blot | 1 |
| decarboxylase assays | carboxylase assays | 2 |
| facs | aas | 2 |
| vitro tme models | vitro eye models | 2 |
| sar assays | nab assays | 2 |
| aldolase assays | hydrolase assays | 3 |
| asthma | asapms | 3 |
| bone models | mouse models | 3 |

Levenshtein distance can be misleading

# Second Metric: Cosine Similarity

- Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space
- It is defined to be equal to the cosine of the angle between the vectors
- “S” is measured in the range 0 to 1
  - Exact match if $S = 1$
  - Approximate match if $1 > S \geq 0.5$
  - No match if $S < 0.5$




https://medium.com/@adriensieg/text-similarities-da019229c894

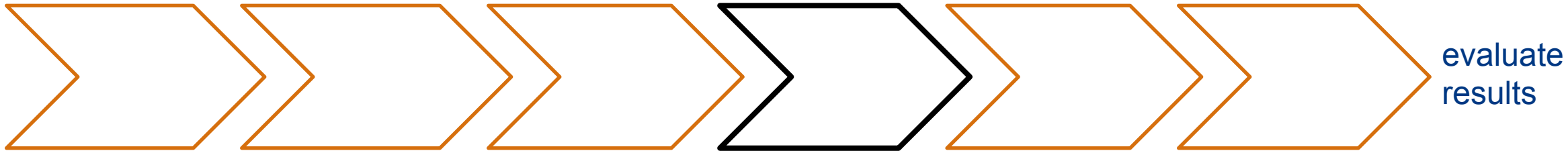
# Natural Language Processing for String Similarity Search

eurofins | DISCOVERY

- The Python Sklearn package contains algorithms for the necessary code-fu
  - Vectorising
  - Natural Language Processing
  - Similarity calculation
  - Ranking
- String similarity was calculated in Python based on article by Dario Radečić

https://towardsdatascience.com/calculating-string-similarity-in-python-276e18a7d33a

# Cosine Similarity: NLP Workflow




https://en.wikipedia.org/wiki/Cosine_similarity

# Matches Ranked by Term Similarity: Excel Output

eurofins | DISCOVERY

Look across

Look Down

| Query Term | Match_1 | CSim_1 | Match_2 | CSim_2 | Match | CSim_ | Match | CSim_ | Match | CSim_ | Match | CSim_ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Systemic lupus erythematosus (S | Systemic Lupus Erythematosus (SLE) Hur | 0.816 | c Lupus Erythematosus (SLE) Hur | 0.816 | Systemic L | 0.816 | Systemic L | 0.756 | Systemic L | 0.632 | Lu Anin | 0.289 |
| Cancer Systems Biology | Systems Biology | 0.816 | ational S | | | | | | | | us | 0.408 |
| Thermal and Acoustic Imaging | Thermal Imaging | 0.816 | Imaging | | | | | | | | | 0.408 |
| Thrombosis Models | Thrombosis Animal Models | 0.816 | odels | | | | | | | | ar | 0.5 |
| Wistar Rat | Wistar Rat Model | 0.816 | KYOTO Rat Model | 0.707 | Rat Models | 0.5 | BDIX/Orllcc | 0.408 | Custom Ra | 0.408 | Kn -In F | 0.408 |
| Zucker Rat | Zucker Rat Model | 0.816 | dels | 0.5 | BDIX/Orllcc | 0.408 | Custom Ra | 0.408 | Knock-In F | 0.408 | Knockout F | 0.408 |
| Inflammatory Pain Models | Acute Inflammatory Pain Animal Models | 0.775 | Inflammatory Pain Animal Models | 0.775 | Custom Pa | 0.577 | Neuropathi | 0.577 | Nociceptiv | 0.577 | Postoperat | 0.577 |
| In Vivo Whole Tissue and Animal I | Animal in vivo Imaging | 0.775 | hole Brain Imaging | 0.671 | Animal Wh | 0.671 | Small Anim | 0.6 | Animal Ima | 0.516 | Animal Tiss | 0.516 |
| Non GLP Large Molecule Bioana | Large Molecule Bioanalysis | 0.775 | olecule Bioanalysis | 0.516 | Bioanalysis | 0.447 | Bead-Base | 0.316 | GLP Auditi | 0.316 | Plasma Bio | 0.316 |
| Non GLP Small Molecule Bioanal | Small Molecule Bioanalysis | 0.775 | olecule Bioanalysis | 0.516 | Bioanalysis | 0.447 | Bead-Base | 0.316 | GLP Auditi | 0.316 | Plasma Bio | 0.316 |
| Amino Acid Analysis (AAA) | Amino Acid Substitution Analysis | 0.75 | ino Acid Analysis | 0.75 | Hydrolyzed | 0.75 | Bile Acid A | 0.577 | Fatty Acid | 0.577 | Nucleic Ac | 0.577 |
| Copy Number Variant Analysis | Gene Copy Number Analysis | 0.75 | Variant Analysis | 0.577 | Genetic Va | 0.577 | 5-Batch Ar | 0.354 | Analysis of | 0.354 | Biomarker | 0.354 |
| Gene Set Enrichment Analyses | Gene Set Enrichment Analysis | 0.75 | notation | 0.354 | Gene Char | 0.354 | Gene Fragi | 0.354 | Gene Knoc | 0.354 | Gene Syntl | 0.354 |
| Antibody Efficacy Testing In Vivo | In vivo Drug Efficacy Testing | 0.75 | Testing | 0.707 | Ex vivo Dru | 0.671 | Antimicrobi | 0.577 | In vivo Antib | 0.577 | In vivo Bios | 0.577 |
| LPS Lung Inflammation Models | Lung Inflammation Animal Models | 0.75 | flammation Models | 0.577 | Inflammatic | 0.577 | Acute Inflar | 0.5 | Chronic Infl | 0.5 | LPS-Induc | 0.5 |
| Rabbit Monoclonal Antibody Dev | Monoclonal Antibody (mAb) Development | 0.75 | Development | 0.707 | Therapeuti | 0.671 | BiTE Antibc | 0.577 | Polyclonal | 0.577 | Monoclona | 0.5 |
| Pulmonary Hypertension Animal M | Pulmonary Hypertension Animal Models | 0.75 | one Animal Model | 0.577 | Diuresis An | 0.577 | Hypertensi | 0.577 | Animal Moc | 0.5 | Animal Moc | 0.5 |
| In Vivo Whole Tissue and Animal I | Animal in vivo Imaging | 0.707 | hole Brain Imaging | 0.612 | Animal Wh | 0.612 | Tissue Ana | 0.577 | Small Anim | 0.548 | AFM Imagir | 0.471 |
| Circulating Antibody Assay Devel | Antibody Development | 0.707 | evelopment | 0.707 | Antibody S | 0.577 | Assay Deve | 0.577 | Biochemic | 0.577 | Biomarker | 0.577 |
| Antibody and Protein Products H | Antibody Products | 0.707 | Protein Sequencing | 0.577 | Membrane | 0.5 | Alginate Pr | 0.354 | Antibody C | 0.354 | Antibody C | 0.354 |
| Autoimmune disease | Autoimmune Disease Animal Models | 0.707 | Models of Disease | 0.408 | Celiac Dise | 0.408 | In vitro Dise | 0.408 | Infectious D | 0.408 | Lyme Dise | 0.408 |
| Autoimmune Models | Autoimmune Disease Animal Models | 0.707 | ental Autoimmune Encephalomye | 0.577 | Experiment | 0.577 | Experiment | 0.577 | Avian Mode | 0.5 | Chicken M | 0.5 |
| BALB/c mouse | BALB/c Inbred Mouse Model | 0.707 | nu/nu Mouse Model | 0.707 | Mouse Moc | 0.5 | Mouse Phe | 0.5 | A/J Mouse | 0.408 | Agouti Mou | 0.408 |
| Sample Bioanalysis | Bioanalysis | 0.707 | ased Bioanalysis | 0.5 | Plasma Bio | 0.5 | Sample Pa | 0.5 | Sample Re | 0.5 | Clinical Sar | 0.408 |
| Chemokine Biomarkers | Biomarkers | 0.707 | Biomarkers | 0.5 | Serum Bior | 0.5 | Pharmacoc | 0.408 | 1,25-Dihyd | 0 | 10X Genom | 0 |
| Cytokine Biomarkers | Biomarkers | 0.707 | Biomarkers | 0.5 | Cytokine A | 0.5 | Serum Bior | 0.5 | Cytokine R | 0.408 | Cytokine S | 0.408 |
| Translational Biomarkers | Biomarkers | 0.707 | Biomarkers | 0.5 | Serum Bior | 0.5 | Pharmacoc | 0.408 | Translation | 0.354 | 1,25-Dihyd | 0 |
| Metastasis Models | Bone Metastasis Animal Models | 0.707 | odels | 0.5 | Chicken M | 0.5 | Insect Mod | 0.5 | Invertebrat | 0.5 | Mammalian | 0.5 |
| CRISPR | CRISPR Bioinformatics | 0.707 | Cloning | 0.707 | CRISPR Sc | 0.707 | CRISPR De | 0.577 | CRISPR sg | 0.577 | Custom CR | 0.577 |
| Custom Antibodies | Custom Affinity Purification of Antibodies | 0.707 | Manufacturing | 0.5 | Custom Mic | 0.5 | Custom Pr | 0.5 | Llama Antib | 0.5 | Custom An | 0.408 |
| Model Development | Custom Cell Model Development | 0.707 | ic Animal Model Development | 0.707 | Algorithm D | 0.5 | Antibody D | 0.5 | App Develc | 0.5 | Assay Deve | 0.5 |
| Pain Models | Custom Pain Animal Models | 0.707 | Neuropathic Pain Animal Models | 0.707 | Nociceptiv | 0.707 | Postoperat | 0.707 | Visceral Pa | 0.707 | Acute Inflar | 0.632 |
| Cytotoxicity | Cytotoxicity Assays | 0.707 | Real-Time Cytotoxicity Assay | 0.577 | Medical De | 0.5 | Natural Kill | 0.447 | 1,25-Dihyd | 0 | 10X Genom | 0 |
| Electroporation Transfection | Electroporation | 0.707 | Transfection | 0.707 | Chemical T | 0.5 | In utero Ele | 0.5 | Microinject | 0.5 | Neuronal E | 0.5 |
| Endocrine Models | Endocrine Disease Animal Models | 0.707 | Avian Models | 0.5 | Chicken M | 0.5 | Insect Mod | 0.5 | Invertebrat | 0.5 | Mammalian | 0.5 |

# Cosine Similarity Approximate Matches Work Fairly Well

eurofins | DISCOVERY

| Query Term | Match_1 | CSim_1 | Match_2 | CSim_2 |
|---|---|---|---|---|
| Acetyltransferase Assays | Acetyltransferase Assays | 1 | Histone Acetyltransferase (HAT) Assays | 0.707 |
| Activity Prediction | Activity Prediction | 1 | Animal Antiplasmin Activity | 0.408 |
| Acute Kidney Failure Models | Acute Kidney Failure Models | 1 | Acute Liver Failure Animal Models | 0.671 |
| wound healing | In vitro Wound Healing Assay | 0.707 | Wound Healing Animal Models | 0.707 |
| In Vivo Efficacy | In vivo Drug Efficacy Testing | 0.707 | Ex vivo Drug Efficacy Testing | 0.632 |
| LCMS Analysis | LC-MS | 0.707 | | |
| Multidrug Resistance | Multidrug Resistance (MDR) Testing | 0.707 | | |
| DNA Mutagenesis | Mutagenesis | 0.707 | | |
| Membrane Permeability (PAMPA) | Parallel Artificial Membrane Permeability A: | 0.707 | | |
| Respiratory Models | Respiratory Disease Animal Models | 0.707 | Avian Models | 0.5 |
| Cardiovascular Toxicology | Toxicology | 0.707 | Computational Toxicology | 0.5 |
| Cell Transfection | Transfection | 0.707 | B Cell Count | 0.5 |
| RNA Interference | Vector Based RNA Interference | 0.707 | Human RNA | 0.5 |
| Human Primary Cell Disease Model | Human Primary Cell Isolation | 0.671 | Animal Models of Disease | 0.516 |
| Small Cell Lung Cancer Models | Lung Cancer Animal Models | 0.671 | Lung Small Cell Xenograft | 0.671 |
| Adherent Cell Culture | 2D Cell Culture | 0.667 | Actinomycete Cell Culture | 0.667 |
| Method Development and Validati | Analytical Method Development | 0.667 | Analytical Method Validation | 0.667 |
| Disseminated Disease Models | Animal Models of Disease | 0.667 | In vitro Disease Models | 0.667 |
| Annexin V Apoptisis Assay | Annexin V Staining Assay | 0.667 | 1,25-Dihydroxyvitamin D Assay | 0.408 |
| Basic Metabolic Profiling | Basic Metabolic Panel | 0.667 | Compound Profiling | 0.408 |
| Brain Tumor Mouse | Brain Tumor Models | 0.667 | Brain Xenograft | 0.408 |
| Caco-2 Cell Line | Caco-2 Cell Lines | 0.667 | Cell Line Authentication | 0.667 |

MATCHES | MATCHES (2)

In the best match column (Match_1)

Out of 523 terms:

- ~20% exact matches
- ~60% approximate matches
- ~20% unresolved

with this particular contributor group

# Where Are We Now, As Users?

eurofins | DISCOVERY

Finding and using the correct ontology is challenging.
Our ultimate goal is to contribute assay data to an ontology that we find helpful.

- Many biomedical ontologies relevant for cellular assays exist
- Excel is excellent for matching exact terms but poorly matches anything beyond that
- Cosine similarity for string matching works well for approximate matches
- People are still needed for curation
  - A key term, phenotypic screening, missing from the list of 5K terms was added
  - Irrelevant terms (12 %) included by the similarity scoring method were removed

# Could We Have Used A Different Ontology Mapping Tool?

eurofins | DISCOVERY

- Short time frame and resources to complete the project
- We have little expertise with ontologies
- Although there are free ontology mapping tools online…
  …the terms we worked with - we think - were not encompassed by any ontology

# Thank You

Aimee Allen
Jennifer Drake
Ellen Berg

# What If…

- Categories & terms were equally understood / meant the same thing to contributors of diverse backgrounds?
- eMarketplaces were educated about ontologies & established adherence to controlled ontologies?
- Terms were harmonized across content platforms beyond eMarketplaces?
  - assay descriptions
  - web content
  - file properties and keyword tags