MetaWise: Extraction And Normalisation Of Toxicologic Pathology Nomenclature From The INHAND Project For Enhanced Search

Jane Reed, Paul Bradley, Lucy Sheppard, Stephanie Berry, Jennifer Feldmann (Instem)

INTRODUCTION

Naming conventions are notoriously ambiguous in the biosciences, which make search and retrieval of data, data analysis, and data sharing difficult. INHAND (International Harmonization of Nomenclature and Diagnostic Criteria for Lesion in Rats and Mice), is a global initiative to create a harmonised toxicologic pathology nomenclature, aiming to standardise the classification of pathological lesions in toxicity studies[1].

One benefit of this initiative will be the ability to use this standardised nomenclature for investigational search and analysis of toxicity data. At Instem, we are enhancing the INHAND nomenclature by mapping all specific pathology terms into a widely-used over-arching biomedical ontology. Using this ontology, it is possible to search for all studies resulting in some degenerative, vascular, inflammatory or other category of pathology, easily and quickly.

This provides a new resource for clinical and pre-clinical scientists, to enable search and analysis across disparate data types.

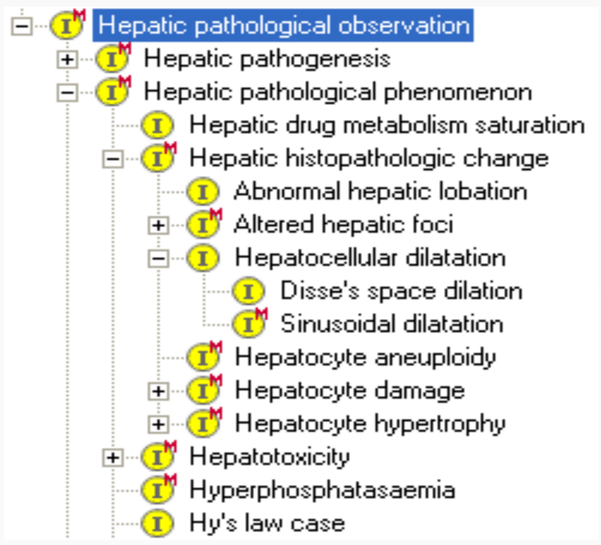


Figure 1: Subset of the hepatic pathological observation branch of the Instem Scientific biomedical observation ontology.

Mapping the INHAND nomenclature to the Instem ontology gives them added richness in the form of synonyms and related terms, and also enables comprehensive search using the taxonomies.

Instem's comprehensive and rich ontology has been used for over a decade by many international pharmaceutical companies.

METHODS

We used Instem's Metawise[tm] to process about 30 INHAND and SSNDC (Standardized System of Nomenclature and Diagnostic Criteria[2]) documents. Metawise identified around 14,000 raw biomedical observation terms. These were manually reviewed after prioritisation based on criteria such as Metawise translation score, term length and frequency within the documents.

Many of these were translated by Metawise to existing biomedical observations within the Instem Scientific biomedical ontology. In addition, Metawise also identified over 2,500 novel biomedical observations, and 1,300 new synonyms for existing biomedical observations; both of these sets of terms were loaded into the appropriate hierarchical nodes in the existing Instem Scientific biomedical ontology (Figure 1).

Nomenclature: a systematic list of names for all known entities within a discipline
Controlled vocabulary: a selected list of names of entities, together with synonyms and related terms (similar to a thesaurus).
Taxonomy: a kind of controlled vocabulary that has a hierarchy (broader term/narrower terms) that enables the user to search up and down a "tree" of related concepts.
Ontology: a representation of knowledge around all concepts within a domain, with attributes and relationships between the various concepts (including synonym, hierarchical , and other relations) that define the domain of knowledge.

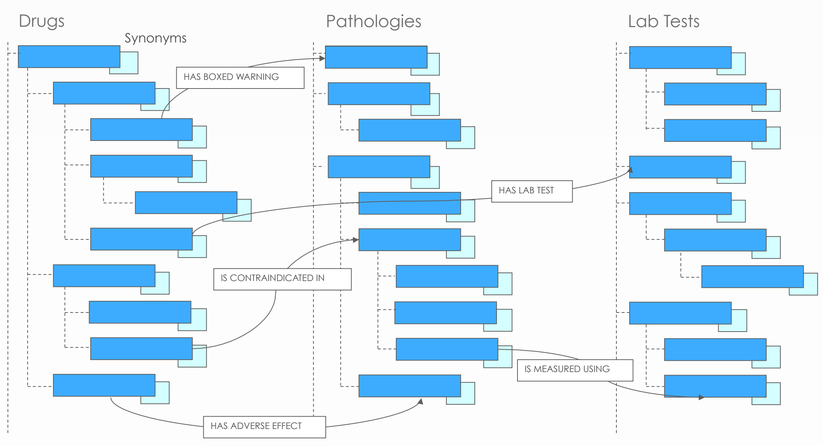
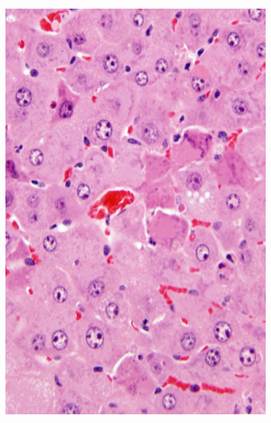


Figure 2: Ontological relations (black arrows with relationships in text boxes), connecting taxonomies of drugs, pathologies and laboratory tests.

Figure 3: Early apoptosis in the liver (courtesy of the Digitized Atlas of Mouse Liver Lesions). Pathologists have many different names for the process shown in the image (liver apoptosis, apoptotic hepatocytes, hepatic apoptosis, apoptotic liver cells), but when scientists need to identify all studies showing apoptosis within liver cells, then translation of all options to a controlled vocabulary is needed.



Metawise

- Identifies, translates and harmonises medically relevant relationships expressed in scientific content
- Provides a tool-kit for performing high-performance concept identification and translation
- Recognises how important concepts are expressed in the real world - aliases, colloquialisms and misspellings
- Based on structure and semantics – greater robustness

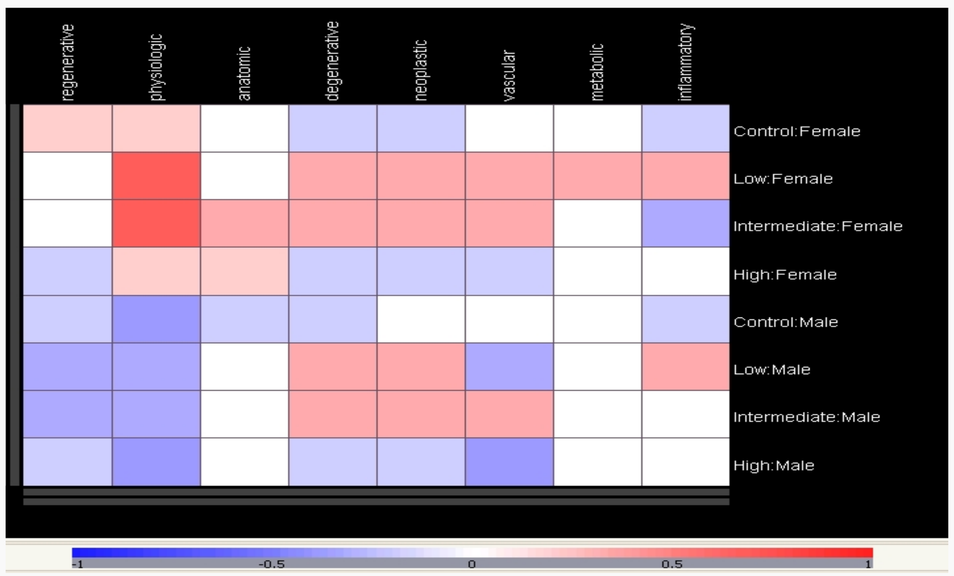
References:

1. International Harmonization of Nomenclature and Diagnostic Criteria for Lesion in Rats and Mice (INHAND). http://www.toxpath.org/inhand.asp

2. Standardized System of Nomenclature and Diagnostic Criteria (SSNDC). http://www.toxpath.org/ssndc.asp

RESULTS

We utilized the focused set of ToxPath concepts within the over-arching biomedical ontology (the "ToxPath Vocab" to visualize non-clinical data sets (Figure 4), and also to search across reports (PubMed abstracts; Figure 4).



Red indicates a higher-than-expected co-occurrence

Blue indicates a lower-than-expected co-occurrence

Figure 4: Using ToxPath Vocab over pre-clinical data, from a 28-day rat study. This OmniViz Comet plot visualizes co-occurrences of selected events – here the different sex and dose groups (vertical labels) vs. the ToxPath high-level taxonomic groupings (e.g. vascular, inflammatory, neoplastic, degenerative). This shows how use of Instem's ToxPath Vocab enables patterns to be seen in complex data.

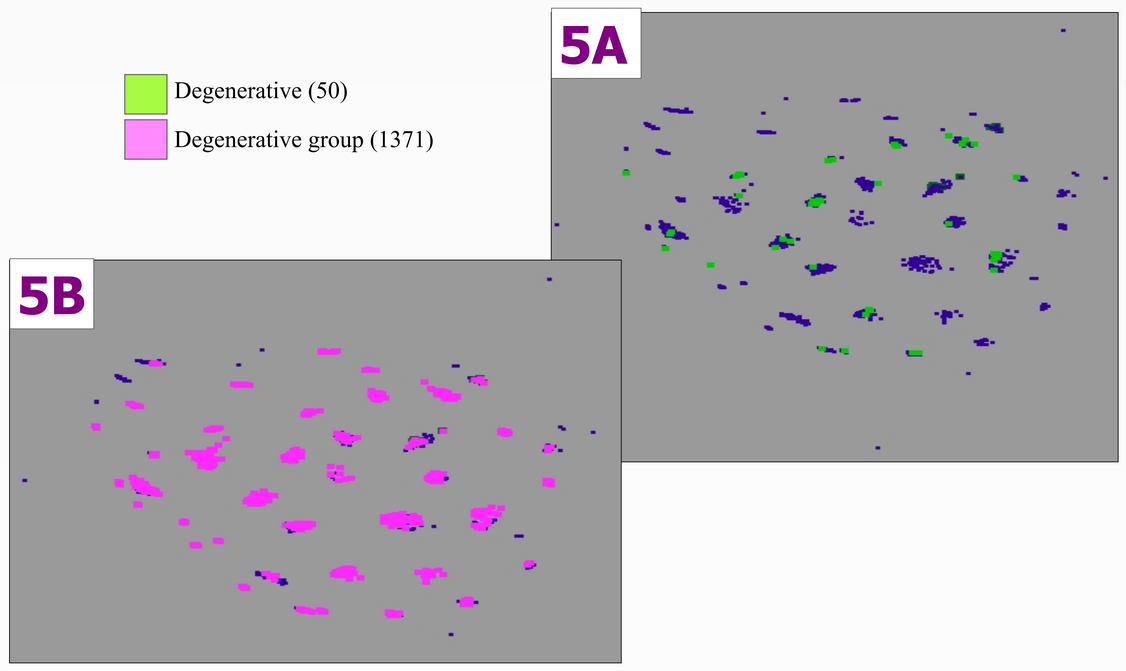


Figure 5: Improved recall for search, using the ToxPath Vocab. The two plots show clustering of abstracts from the journal Toxicologic Pathology (2.6k abstracts; 1983-2012), which have been processed using the ToxPath Vocab.

5A shows highlighting of those abstracts after searching with the term "degeneration", 5B shows the increased recall using the ToxPath Vocab that relate to degeneration (from 50 to 1371 documents). 5C shows mark-up of one of the abstracts with the range of degenerative terms highlighted in yellow; mark-up of non-degeneration terms has been hidden.

CONCLUSIONS

The tremendous growth in biological data demands the use of controlled vocabularies and ontologies, for consistent representation of the information. Harmonisation of such knowledge facilitates comparisons between different datasets and better communication of the knowledge

Application of Instem's Metawise for mark-up and translation creates a consistent metadata layer over the pathology documents, enabling high-level search for histopathology terms.

By extracting and incorporating the INHAND nomenclature we can enable high-level search for histopathology terms that might be indicative of general pathological processes (e.g. inflammation, degeneration and regeneration).

This work provides a substrate for the further development of an improved biomedical observation ontology, spanning both the pre-clinical (e.g. microhistopathology terms etc.) and the clinical (e.g. human disease terms).

Instem
Information Solutions For Life