PHUSE US CSS 2023: PP30

From FAIR Data to FAIR Knowledge Sharing

Dragomir Draganov[1], Maddalena Marchesi[2], Olivier Roche[3], Stefano Gaudio[1], Rajalaxmi S.[4], Raja T. Ramesh[4]

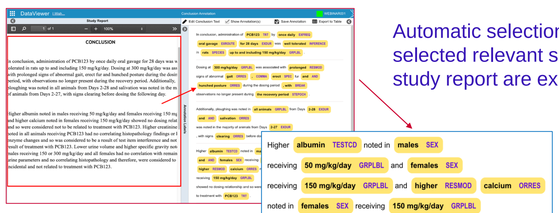
[1]pRED Pharmaceutical Sciences, [2]Product Development, [3]pRED Data & Analytics, F. Hoffmann-La Roche Ltd., [4]PointCross Life Sciences Inc.

1 – Introduction

Previously, we reported SEND datasets warehousing and exploration in the Safety Data Integration (**SDI**) database (Draganov et al., CSS 2020, PP08). In this paper we describe the contextualization of the SEND data with conclusions and interpretation from the study reports and/or subject matter experts (SMEs), e.g., toxicologists, pathologists, study monitors. The information extracted from the study report can be linked to the Tables and Figures generated from the data as collected and/or to subject level data. The conversion of the annotations into structured columnar data allows the creation of informative visualizations for a study and cross-study explorations and easy search within the database for treatment-related and adverse findings. In essence, this this is our application instance of the "SR"-domain proposed by Drew et al. (CSS 2019, PP07) with some additional information captured and links/tags to Tables, Figures & Listing (TFL) objects in SDI, but also expanded beyond the interpretation of a single study findings and observations.

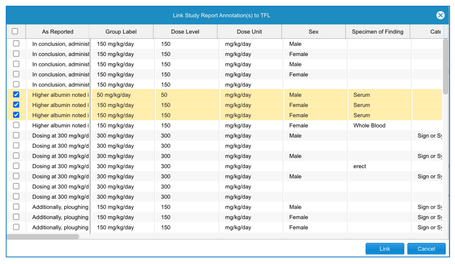
One of the main goals of setting up/maintaining a data base for SEND data is the reusability of the latter according to the FAIR principles described by Wilkinson at al. (2016). There appear to be a misconception that the high level of standardization of the SEND data coupled with their local [database] findability/accessibility ensures their FAIR status, thus overlooking the actual interoperability of the data on an organizational level. To increase the FAIRness of the data in SDI (which is rather a process than a one-time campaign), we performed a terminology harmonization with Roche Terminology System (RTS) and pursue integration with Roche's **F**AIR **I**n vivo Data **SH**aring (FISH) platform. FISH provides global unique, persistent and resolvable identifiers (GUPRIs) for study, animal, treatment and biospecimen registration, alongside data models for the respective domains and an overarching semantic data layer. SDI is connected bidirectionally through APIs with FISH registrations systems as they become available. This provides two major benefits. First, this makes [meta]data in SDI globally (Organization-wide) findable and accessible (based on the business case). Second, it provides an opportunity for linking trough a knowledge graph [meta]data presentation of the low-dimensional data in SDI (most of the endpoints in the safety-relevant studies are low dimensional) to study data not covered by the current SEND standard and/or are high-dimensional endpoints such as toxicogenomics, bulk or single-cell RNA sequencing, spatial proteomics, digital pathology images etc.

2 – Data / TFL Annotations Contextualize the Subject Level SEND Data



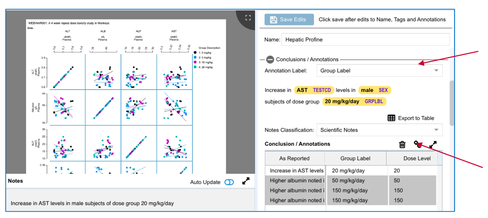
Automatic selection of the Conclusion section plus any other user-selected relevant sections, sentences and text strings from the study report are exported to the 'Conclusion Annotation' tool.

Text is parsed using NLP model based on spaCy's entity recognition (Honnibal and Montani, 2017). Any additional tagging or corrections are then applied manually to complete the annotation process.



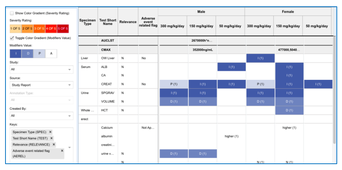
Annotated text for each relevant finding/observation capturing group, sex and subject count is standardized semi-automatically and enabled for editing before saving with the refence to the document page number, created by and updated by, hence traceable to the source.

The summary metadata is modelled to store 48 variables such as relevance to treatment, adversity, reversibility, [statistical] significance etc.

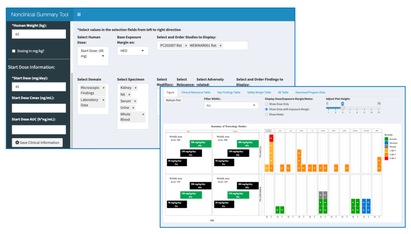


SMEs' annotations of TFL Objects (Tables or Figures) can be converted to tabular standardized output similar to the report annotations.

Report annotations are available as a for linking to a particular TFL Object.



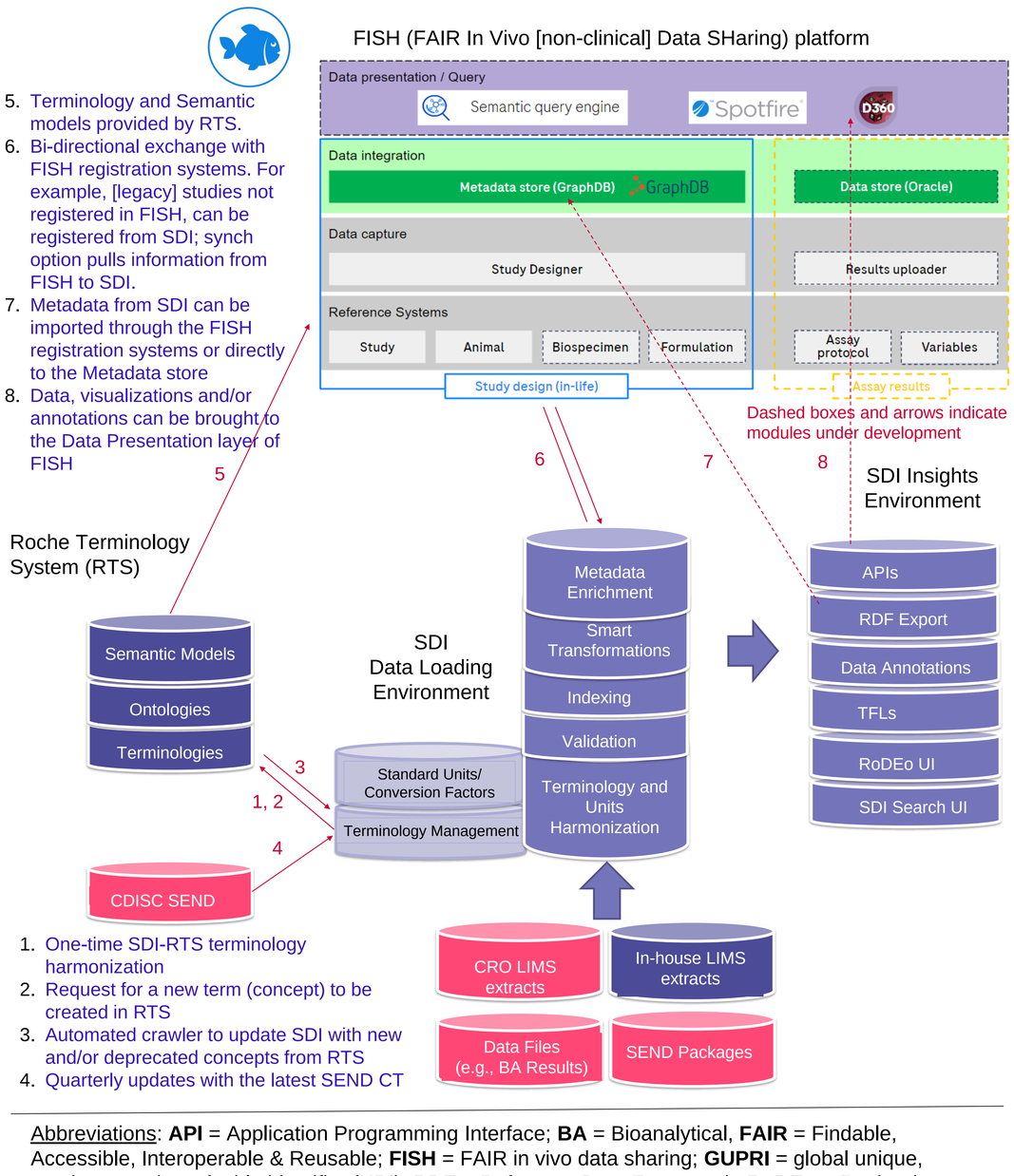
Structured summary information is available for the visualization within a study and across studies.



Cross-study annotations can be exported as a tabular file and imported into ToxSummary R shiny app (Ali and Snyder, 2023) which has been adapted to accept the data from the above standard model table. The tool is also extended to filter the rows based on the domain of findings, relevance to the drug and adversely related for visualizing the significant findings.

The stacked bar graphs, expanded for display by sex, are coloured for severity (1 to 3 for clinical observations [green-blue] or 1 to 5 for macro/microscopic findings [yellow-red]).

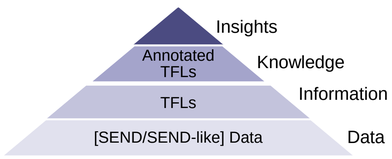
3 – SDI and FISH Integration Provides Global Access to SDI Data and Opportunity to Link to Other Data, e.g., High-Dimensional Data from *In-vivo* Studies



1. One-time SDI-RTS terminology harmonization
2. Request for a new term (concept) to be created in RTS
3. Automated crawler to update SDI with new and/or deprecated concepts from RTS
4. Quarterly updates with the latest SEND CT

Abbreviations: **API** = Application Programming Interface; **BA** = Bioanalytical, **FAIR** = Findable, Accessible, Interoperable & Reusable; **FISH** = FAIR in vivo data sharing; **GUPRI** = global unique, persistent and resolvable identifier (URI); **RDF** = Reference Data Framework; **RoDEo** = Roche data exploration UI; **RTS** = Roche Terminology System; **SDI** = Safety Data Integration; **SME** = Subject Matter Expert; **TFL** = Tables, Figures & Listings; **UI** = User Interface; **URI** = Unique Resource Identifier.

4 – Conclusions



Global access to data in SDI via APIs and metadata through the FISH platform brings the non-clinical safety data to a true FAIR state on an organizational level, which we consider a prerequisite for an efficient and meaningful integration with other types of data (in silico, in vitro, clinical).

Capturing prospectively institutional knowledge in a standardized way (from a study report for a study or IND, NDA, BLA for across studies) and linking it to the subject level data in SDI will facilitate forward and reverse translatability of the non-clinical safety data and may justify the time and effort of gathering this information retrospectively.

5 – References

1. Draganov D. et al.: Taking full advantage of the SEND data potential. PHUSE US CSS 2020: PP08.
2. Drew P, et al.: Consolidating Study Outcomes in a Standardised, SEND-Compatible Structure. PHUSE US CSS 2019: PP07.
3. Wilkinson M. at al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data volume 3, Article number: 160018, 2016.
4. Honnibal M. and Montani I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. Sentometrics Research, 2017.
5. Ali Y and Snyder K: toxSummary: Visualize and Summarize Repeat-Dose Toxicology Study Results. R package version 0.0.0.9000. https://github.com/phuse-org/toxSummary. 2023.

6 – Acknowledgments

SDI is supported by Xbiom 3.3 software developed by PointCross Life Sciences Inc. This collaboration was sponsored by internal Roche pREDi grants - From FAIR data to FAIR Knowledge (2021) and Roche SDI – FISH Integration for SRS, ARS, and SD Modules (2022).
The authors would like to acknowledge the ideas, feedback and contribution of many colleagues at both companies during the development, testing and employment of the work presented in this paper.