OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

Case Study

Data curation and creation of pre-reasoned datasets and searching [**DataCure**]

Table of Contents

SUMMARY

DataCure establishes a process for data curation and annotation that makes use of APIs (eliminating the need for manual file sharing) and semantic annotations for a more systematic and reproducible data curation workflow. In this case study, users are provided with capabilities to allow access to different OpenRiskNet data sources and target specific entries in an automated fashion for the purpose of identifying data and metadata associated with a chemical or other endpoint of interest. The datasets can be curated using an OpenRiskNet services developed for this case study and re-submitted to the data source. Text mining facilities and workflows are also included for the purposes of data searching, extraction and annotation.

A first step in this process was to define APIs and provide the semantic annotation for selected databases (e.g. diXa, FDA datasets, ToxCast and ChEMBL). During the preparation for these use cases, it became clear that the existing ontologies do not cover all requirements of the semantic interoperability layer. Therefore, ontology development and design of the annotation process as an online or an offline/preprocessing step form an ancillary part of this case study.

1. **What we want to achieve?**

   Establish a process for data curation and annotation that makes use of APIs and semantic annotations for a more systematic and reproducible data curation workflow. The development of semantic annotations and API definition for selected databases are also desired.

2. **What we want to deliver?**

   The aim is deliver curated and annotated datasets for OpenRiskNet service users as well as preparation of and development of tools that can allow users perform their own data curation.

3. **What is the solution and how we implement?**

   Developing resources that can make use of APIs as much as possible and eliminate the need for manual file sharing. In addition workflows that provide examples of useful data for toxicogenomic data analysis will be developed.

Expected deliverables:

- APIs
- Data extraction and curation workflows
- Data annotation

Data Source -> API development and annotation (if needed) -> Data Retrieval -> Data quality control (inspect, clean, filter) -> Pre-processed data (Cleaned data) -> merge with other data if needed

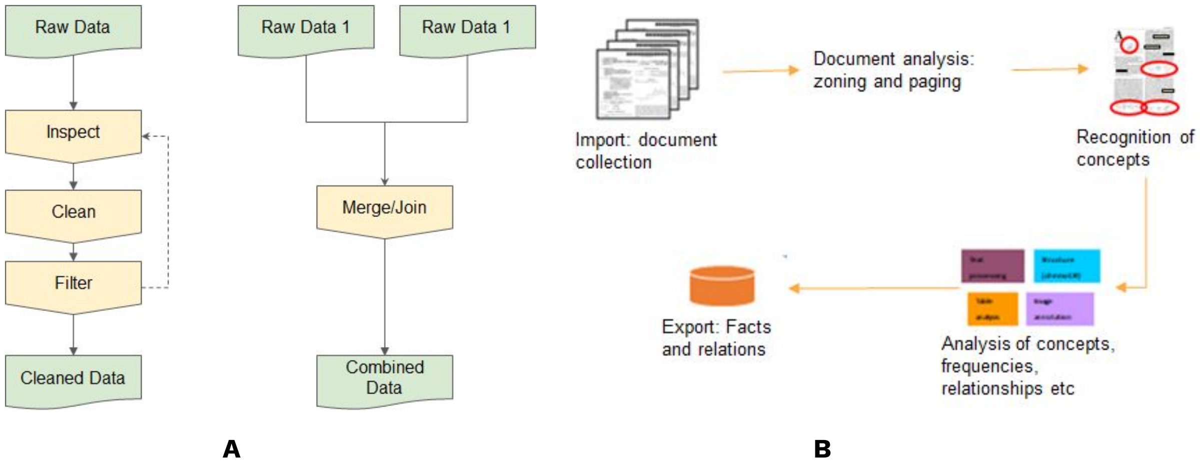
Raw Data
Inspect
Clean
Filter
Cleaned Data
Raw Data 1
Raw Data 1
Merge/Join
Combined Data
Import: document collection
Document analysis: zoning and paging
Recognition of concepts
Analysis of concepts, frequencies, relationships etc
Export: Facts and relations
A
B

Figure 1. A: Raw data curation workflows; B: Text mining workflow

DESCRIPTION

Implementation team

CS leader	Team
Noffisat Oki (EwC)	Thomas Exner (EwC), Tim Dudgeon (IM), Danyel Djennen (UM), Marc Jacobs (Fraunhofer), UoB, NTUA

Case Study objective

- This case study serves as the entry point of curation of all data sources to be used by the remaining use cases;
- Deliver curated and annotated datasets for OpenRiskNet service users as well as preparation and development of tools and workflows that provide examples of useful toxicogenomic data analysis methods that allow users to perform their own data curation and analysis;
- Semantic annotation and API definition for the selected databases will also be carried out in this use case.

DataCure covers the identification of chemical/substance of concern and collection of associated existing metadata. The steps in Tier 0 of the selected risk assessment framework [1]) guide the data retrieval whenever applicable:

- Identification of molecular structure (if required);
- Collection of support data;
- Identification of analogues / suitability assessment and existing data;
- Identification/development of mapping identifiers.

DEVELOPMENT

Use Cases Associated

This case study is associated with UC1 - Merge existing data by a common structure identifier, where a user searches for existing assay information, selects the desired information, and merges the results based on a unique structure identifier. Specifically, the steps to achieve different objectives of the DataCure, include:

- The user identifies and visualises the molecular structure:
    1. Generation of molecular identifiers for database search
    2. Searching all databases
    3. Data curation
    4. Tabular representation
    5. Visualisation
- The user collects supporting data:
    1. Provide data access scheme using the interoperability layer
    2. Access selected databases or flat files in a directory
    3. Query to ontology metadata service and select ontologies, which should be used for annotation
    4. Annotate and index all data sets using text mining extraction infrastructure
    5. Passing to ontology reasoning infrastructure
    6. Generate database of pre-reasoned dataset (semantic integration)
    7. Allow for manual curation
- The user identifies chemical analogues:
    1. Inventory of molecules (commercially available or listed in databases)
    2. Generate list of chemically similar compounds
    3. Collect data of similar compounds

Databases and tools

The following set of data and tools are proposed to be used and exploited within the DataCure:

- Chemical, physchem, toxicological and omics databases: RDKit, CDK, Chemical Identifier Resolver (NIH), PubChem, registries (e.g. ECHA, INCI), EdelweissData Explorer (EwC), Liver Toxicology Knowledge Base (LTKB)
- Ontology/terminology/annotation: SCAIView/JProMiner/BELIEF (Fraunhofer), openBEL
- Literature databases: Pubmed and Toxplanet

Service integration

A set of physical-chemical properties, prediction, workflows, and ontology services will be integrated including the SCAIView, Jaqpot, Conformal prediction, and Jupyter notebooks.

Technical implementation

There are several steps involved in the technical implementation of this case study as listed below. All workflows mentioned here are made available to the public through the use of workflows written in scripting languages (such as R or Python) and prepared in Jupyter or Squonk notebooks. These workflows and notebooks and stored on the publicly accessible OpenRiskNet github.

1. Data sources: The EdelweissData Explorer (further described below) will serve as one of the main data provisioning tool for this case study and others in the project. There are also other sources that are used such as Pubmed and Toxplanet which are repositories primarily storing literature and literature based data
2. Data Extraction: The retrieval of data from the EdelweissData Explorer and all other resources will be done through the use of API calls. Workflows with examples of various forms of data extraction using these APIs are documented and run from Jupyter notebooks. These extraction may also involve additional text mining using the SCAIView tool also integrated into the notebooks.
3. Data Searching: Workflows that employ text mining capabilities are used for searching for specific data and refinement and curation of the data extracted from these sources.
4. Data curation and reasoning: This will be done through the provision of workflows stored in Jupyter or Squonk notebooks. These workflows will cover components such as extraction of specific data and merging of datasets for downstream analysis purposes.
5. Resubmission to data source: This curated datasets may then be resubmitted to the EdelweissData Explorer to be used by others through API access.

Description of tools mentioned in technical implementation

1. EdelweissData Explorer: This platform is a web based data explorer tool that data can be uploaded to and extracted from (using csv files or APIs) to show a nice interface to the data. It gives users the ability to filter, search and extract the current user operation through URLs from the user query. The URL can be used to query for exactly the currently visible dataset directly from EdelweissData APIs in a coding environment.
2. Jupyter Notebooks: This is an open-source web application that provides an interactive programming environment to create, share, and collaborate on code and other documents.
3. Squonk Computational Notebooks: This is an open-source web application provided by IM that is somewhat similar in concept to Jupyter, but targeted at the scientist rather than the programmer.
4. SCAIView: This is an text-mining and information retrieval tool that uses semantic and ontological searches to extract relevant information from a variety of data sources. It can work online or offline to access both publicly available or proprietary data sources available in various formats.

OUTCOMES

Outcome from this case study provide a real problem scenario and use a set of toxicogenomic data along with accompanying metadata to illustrate the processes described above. A webinar demonstrating the steps mentioned in the technical implementation steps was given on the 18th of March 2018.

In this demonstration, the case study participants introduced attendees to the OpenRiskNet data handling and curation process. This included methods for data access, upload, and extraction for further downstream analysis as described in the technical implementation steps.

Specific examples demonstrated during the webinar include:

- A workflow for transcriptomics data extraction and metadata annotation from data stored in EdelweissData Explorer;
- A text mining workflow for metadata extraction for carcinogenicity predictions;
- Demonstration of extraction and curation of data for liver toxicity modeling using data from the Liver toxicity knowledgebase (LTKB) a public database.

The Jupyter notebook workflows prepared for the webinar demonstrations are available here (https://github.com/OpenRiskNet/notebooks/tree/master/DataCure/LTKB) in the OpenRiskNet github.

REFERENCES

1. Berggren E, et al. Ab initio chemical safety assessment: A workflow based on exposure considerations and non-animal methods. Computational Toxicology. Elsevier; 2017;4: 31–44.