A semi-automated pipeline for metadata cleanup

Fawaz Dabbaghie¹, Felipe Albrecht^{1,2}, Markus List¹

E-mail: markus.list@mpi-inf.mpg.de

- 1. Max Planck Institute for Informatics, Saarland Informatics Campus, 66123 Saarbrücken, Germany
- 2. Graduate School of Computer Science, Saarland Informatics Campus, 66123 Saarbrücken, Germany



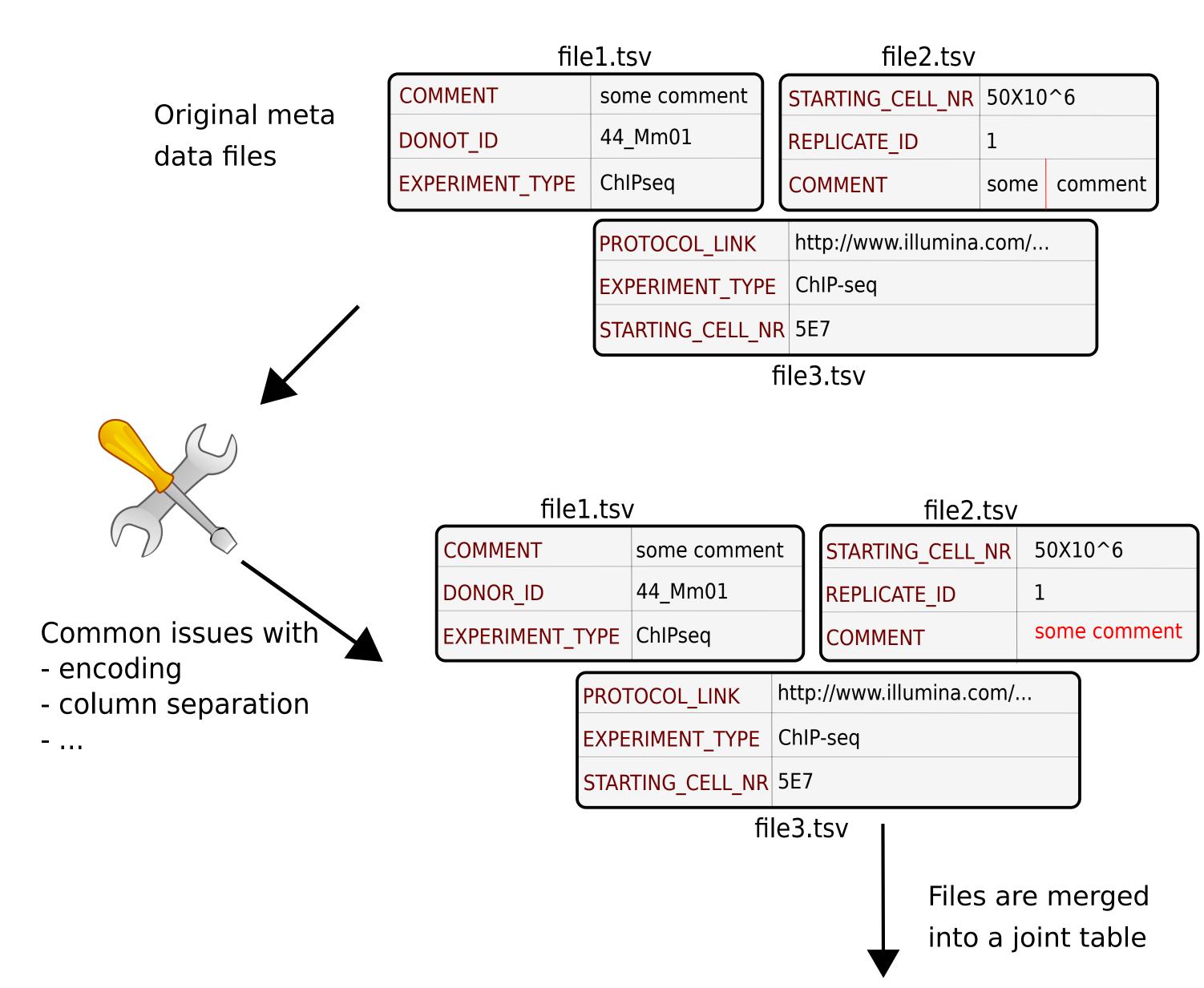
Abstract

Metadata, which include details on samples, experimental protocols or data processing, are a cornerstone of reproducible research. However, keeping metadata in a structured and consistent form is challenging, in particular in large consortia that often span various labs with different conventions. This poses a serious problem, e.g. due to inconsistent use of terms (e.g. ChIPseq vs ChIP-seq) as well as formatting and encoding problems. Typically, these issues are solved in a laborious, manual fashion. As a result, analysis will often focus on a small set of terms, often neglecting potential confounding factors such as software versions, laboratory protocols or devices used.

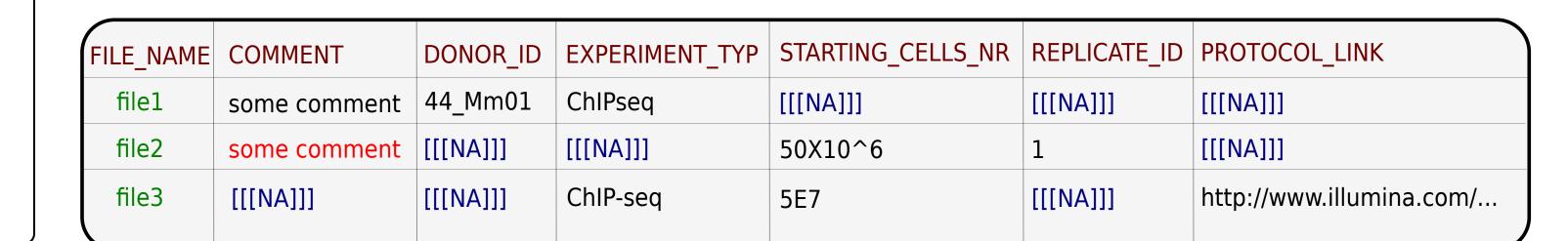
In order to tackle this issue, we implemented a semi-automated pipeline with the following steps: (i) common formatting issues in existing metadata files are addressed; (ii) the user may interact with a unified version of the metadata through OpenRefine, a tool that facilitates interactive and user-friendly mass editing operations. This step yields a set of editing rules that are extracted for future operations; (iii) based on the cleaned up metadata, a controlled vocabulary is generated and complemented by user-defined regular expressions.

Completing this procedure once allows us to address previously encountered issues automatically when metadata for additional samples or experiments become available. Our pipeline, which consists of a series of freely available and well-documented python scripts (https://github.molgen.mpg.de/DEEP/tidy meta data/), allows data analysts to save substantial time. Moreover, the cleaned up metadata allows for potential confounding factors to be considered without additional effort. We successfully applied this strategy within the German Epigenome consortium DEEP and envision that our effort are useful to other IHEC members and beyond.

(i) Preparing and fixing metadata files

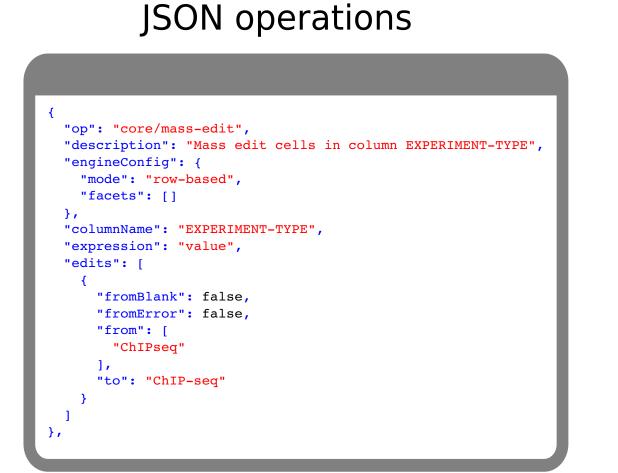


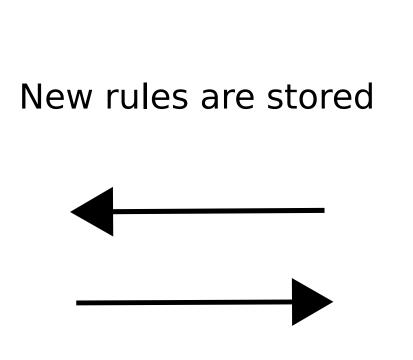
(ii) Batch editing with OpenRefine





available online





Existing rules are automatically applied

to the table of merged metadata



file1.tsv

EXPERIMENT_TYPE ChIP-seq

DONOR ID

44_Mm01

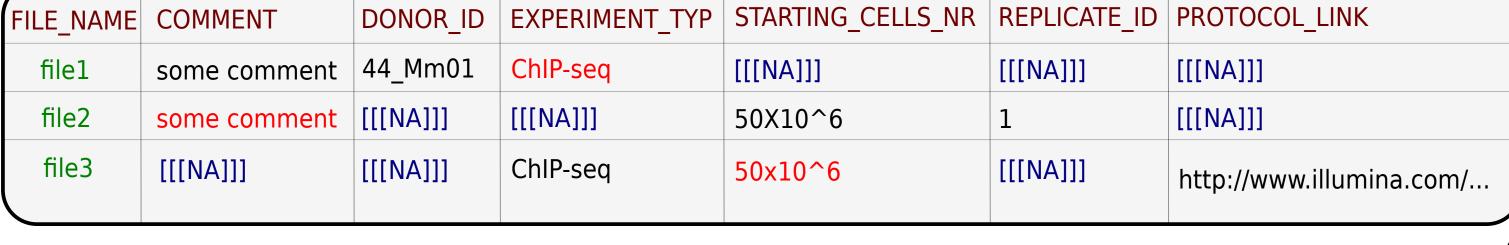
EXPERIMENT_TYPE | ChIP-seq

STARTING_CELL_NR 50x10^6

User defines editing rules interactively

cleaned up files

are produced



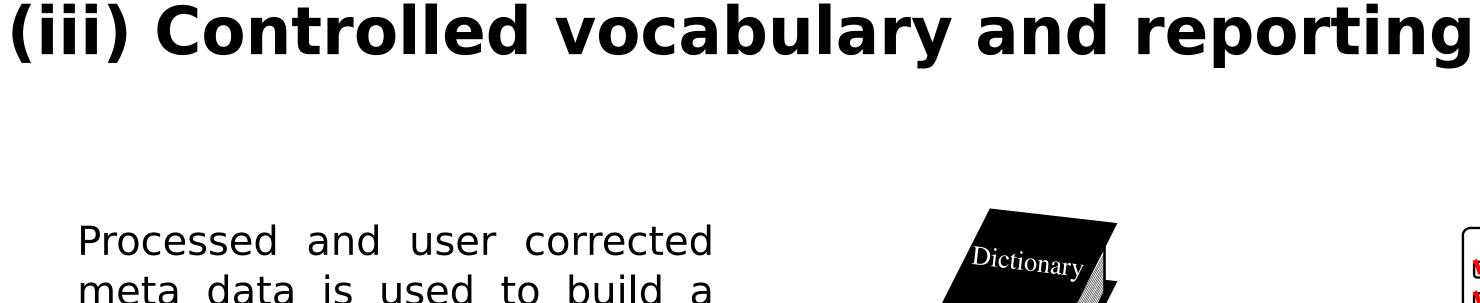
file2.tsv

STARTING CELL NR 50X10^6

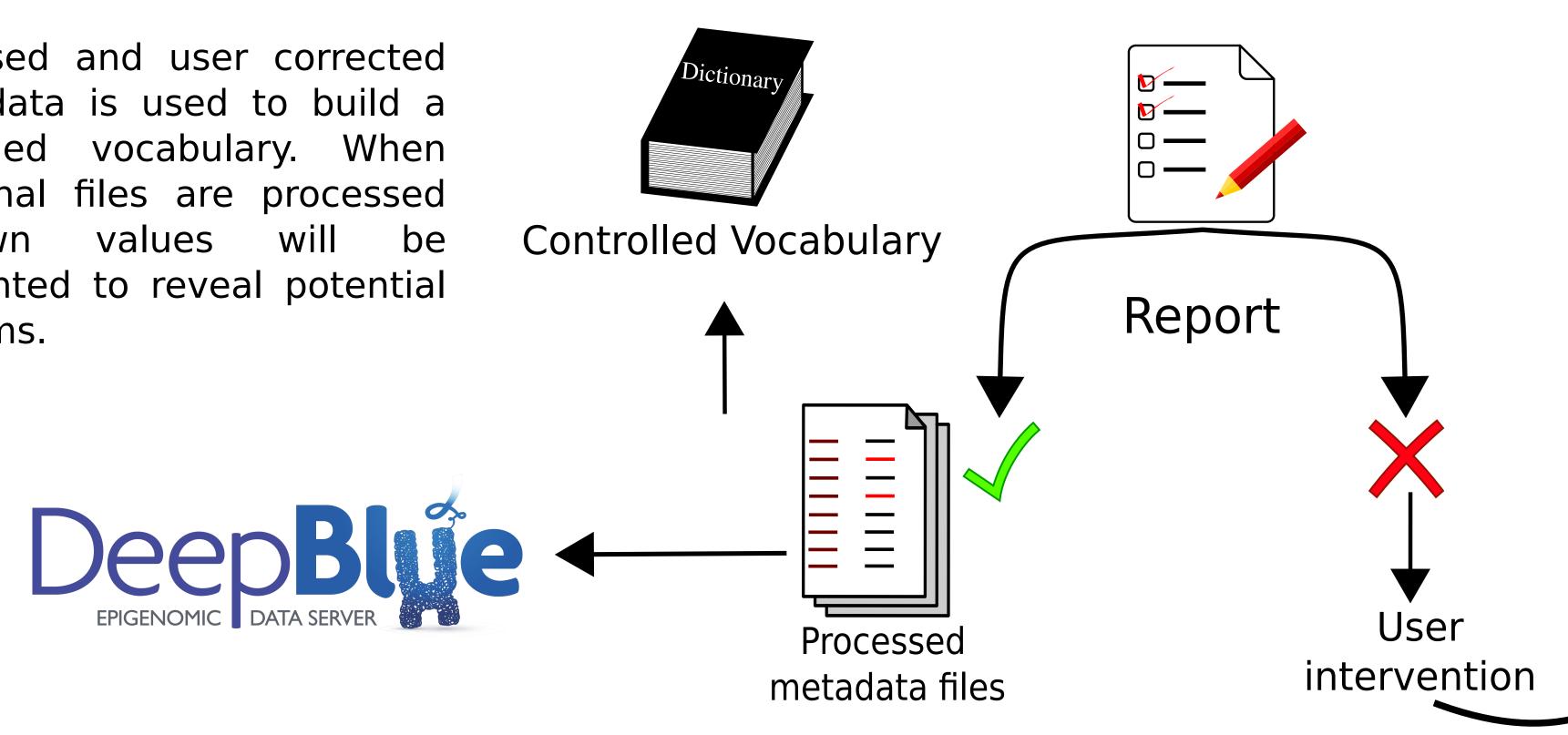
REPLICATE ID

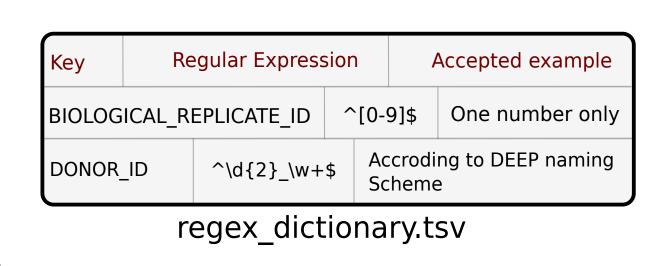
PROTOCOL_LINK http://www.illumina.com/...

file3.tsv



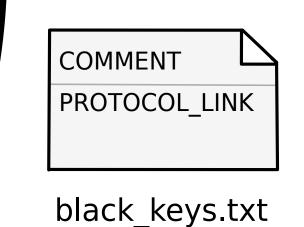
meta data is used to build a controlled vocabulary. When additional files are processed values will unknown highlighted to reveal potential problems.





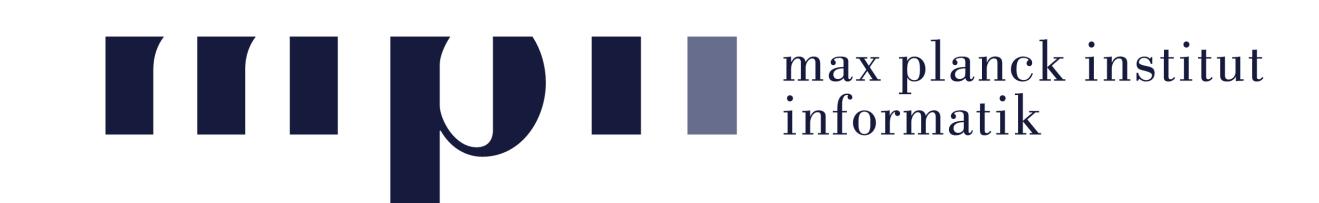
some commen

Users may specify regular expressions to control particular keys that specify e.g. a volume or a concentration.



Users add additional information about blacklisted keys that can be ignored.







GEFÖRDERT VOM