SATORI

"Seeing into one's true nature"

悟り

A System for Ontology-Guided Visual Exploration of Biomedical Data Repositories

Fritz Lekschas[1,2*] and Nils Gehlenborg[1†]

1 Department of Biomedical Informatics, Harvard Medical School
2 Harvard John A. Paulson School of Engineering and Applied Sciences
* lekschas@g.harvard.edu † nils@hms.harvard.edu

Acknowledgement and Funding: We would like to thank the participants in our field studies and the members of the Refinery Platform team. This work was funded by the National Institutes of Health (R00 HG007583) and the HarvardStem Cell Institute.

SATORI is an ontology-guided visual exploration system for data repositories, which combines powerful metadata search with a treemap and a node-link diagram that visualize the repository structure, provide context to retrieved data sets, and serve as an interface to drive semantic querying and exploration, and thereby support the information foraging loop. SATORI is web-based, open-source, and integrated in the Refinery-Platform—an application for biomedical data management, analysis, and visualization.

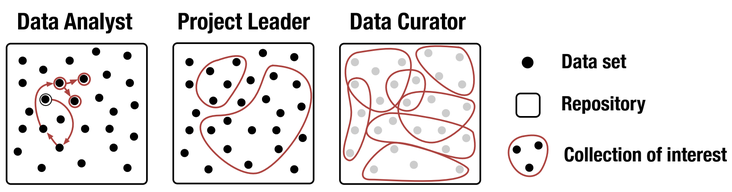

Fig. 2: Exploration behavior of different user roles. Data analysts aim at locating specific data sets. Project leaders focus on collections of data sets and the bigger picture. And data curators are primarily interested in the overall annotation hierarchy.

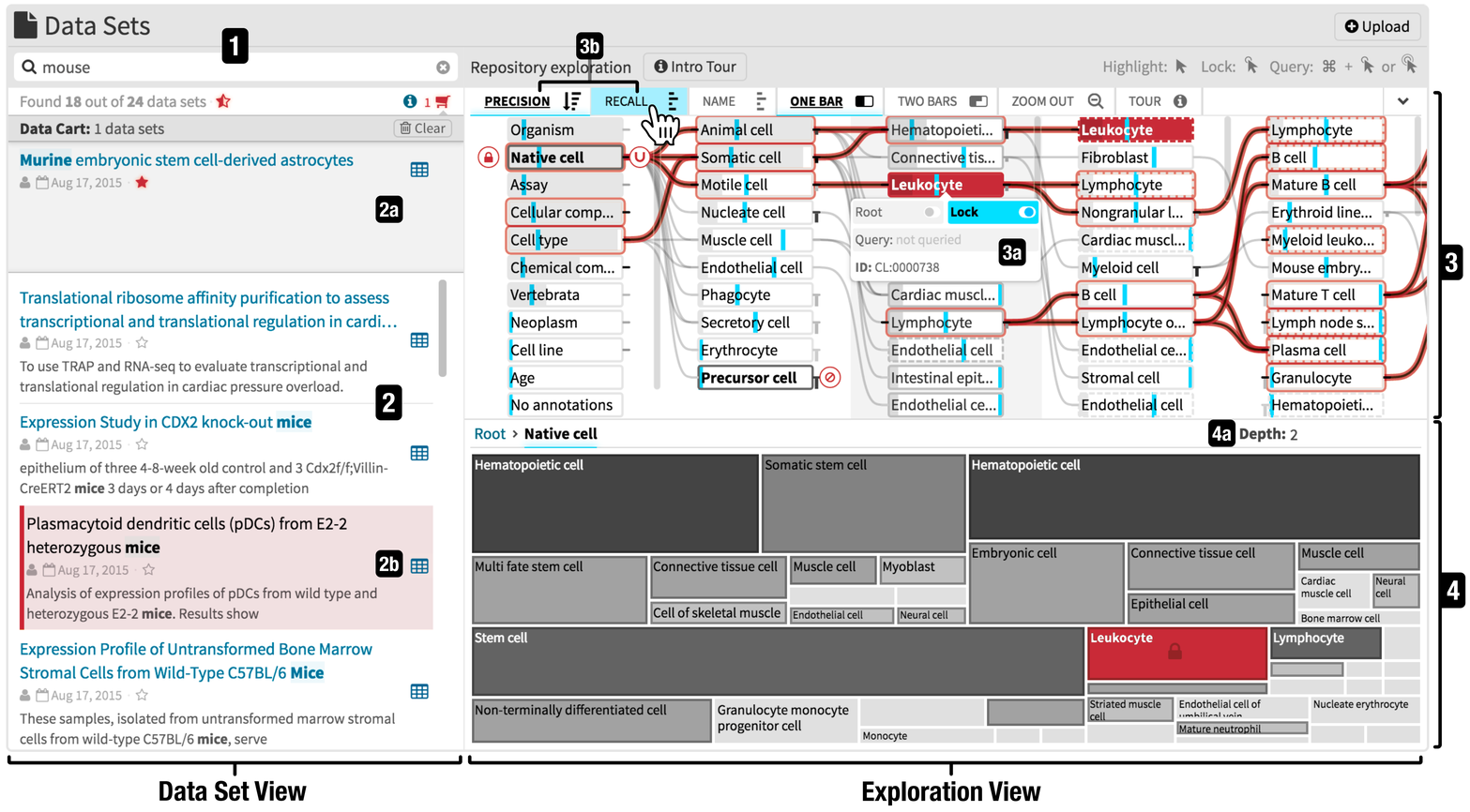

Fig. 1: SATORI's interface: illustrating a query for *native cells* excluding *precursor cells* combined with a synonym keyword search for *mouse* and highlighting *Leukocyte*-related data sets. The data set view includes the search interface (1) and the list of retrieved data sets (2) showing the data cart (2a) and a data set related to *Leukocyte*. The exploration view is composed of a node-link diagram (3) and a treemap (4), which show the (relative) precision and recall of annotation terms among the retrieved data sets. The highlighting of *Leukocyte* is enabled through the node context menu (3a). *Leukocyte* is highlighted in (4) since the visible depth (4a) has been increased to two, i.e., the treemap shows all inner nodes of depth two and all leaves up until depth two. Mousing over the recall button (3b) highlights recall of all terms in (3) while the bar still displays precision.

Why?

Biomedical repositories are growing rapidly and provide scientists with tremendous opportunities to re-use data. In order to exploit published data sets efficiently, it is crucial to understand the content of repositories and to discover data relevant to a question of interest. These are challenging tasks, as most repositories currently only support finding data sets through text-based search of metadata and in some cases also through metadata-based browsing. To address this, we conducted a task analysis through semi-structured interviews with 8 PhD-level domain experts and identified 3 distinct user roles (Figure 2).

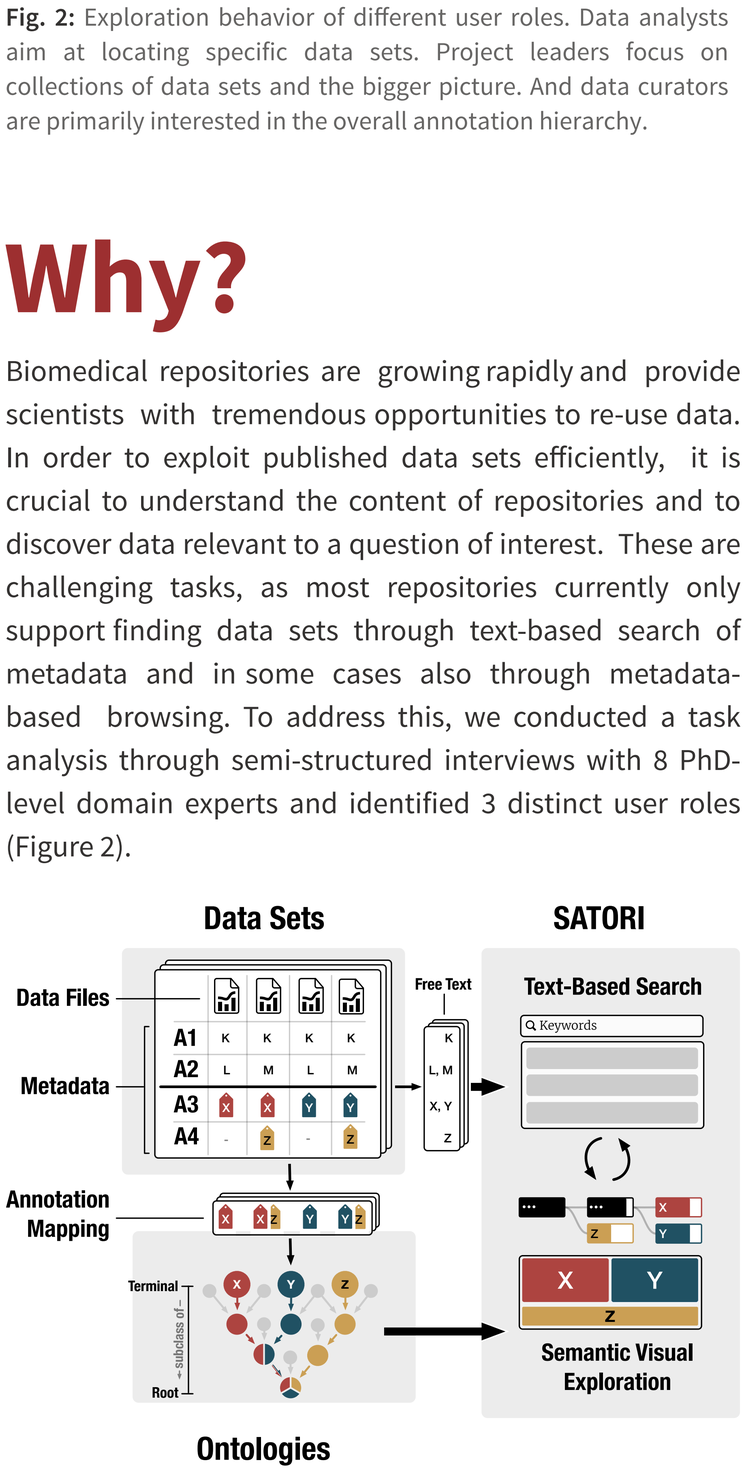

Fig. 3: SATORI system and data model.

What?

Biological data sets consists of experimental data and metadata describing the studies, properties of the analyzed biological samples, and attributes of individual data files.

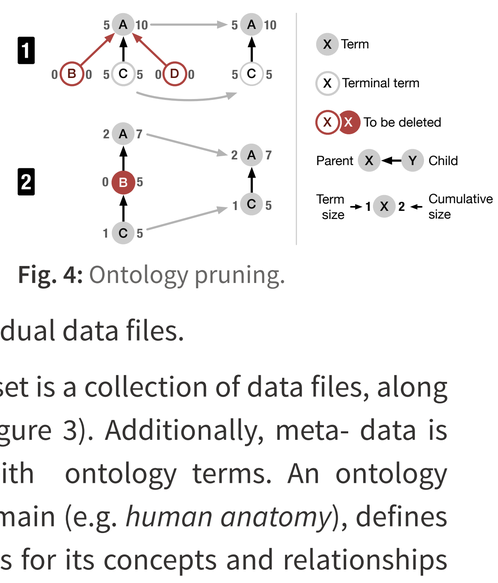

Fig. 4: Ontology pruning.

In this context, a data set is a collection of data files, along with the metadata (Figure 3). Additionally, meta- data is partially annotated with ontology terms. An ontology describes a certain domain (e.g. *human anatomy*), defines controlled vocabularies for its concepts and relationships (e.g., *kidney* and *is-part-of*) and relates concepts with each other (e.g., *nephron is-part-of kidney*). By means of ontology terms, sets of annotated data sets can be classified hierarchically.

SATORI extracts free-text and ontologically annotated metadata (Figure 3). The free-text metadata is indexed in a text-based search system. Additionally, data set-related ontology classes are parsed and visualized to provide semantic context to data sets. Since SATORI's goal is to support exploration rather than to visualize ontologies themselves, only a relevant subtree of the ontologies is shown, i.e., effectively enforcing a strict containment hierarchy (Figure 4).

How?

SATORI is composed of two main interlinked views: the data set view and the exploration view (Figure 1). In the treemap an ontology term is illustrated by a rectangle. The area of the rectangle visualizes the size of the term relative to its sibling terms and the color indicates the distance to the farthest child term. The farther away this child term is, the darker is the color. The node-link diagram represents ontology terms as nodes and links shown parent and child terms (Figure 5). Additionally, the diagram visualizes the precision and recall (Figure 6) for each term given the currently retrieved data sets. In this context, precision is useful to understand how frequently a term is used for annotation in the retrieved set of data sets and recall provides a notion of information scent by indicating if there are more data sets annotated with this term. Finally, the exploration view acts as a semantic query interface and lets users filter down collections of data sets via ontology term-based Boolean queries.

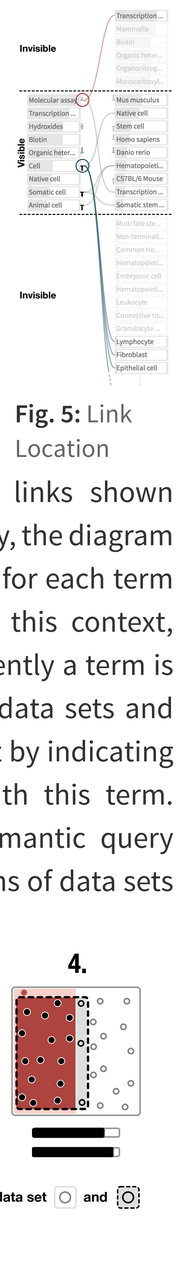

Fig. 5: Link Location

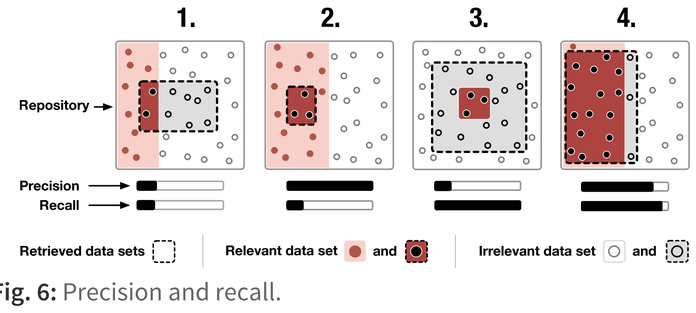

Fig. 6: Precision and recall.

HARVARD MEDICAL SCHOOL

HARVARD John A. Paulson School of Engineering and Applied Sciences

SATORI

Semantic **A**nno**T**ations and **O**ntological **R**elations **I**nterface

http://satori.refinery-platform.org
http://gehlenborglab.org