bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

Fritz Lekschas[1] and Nils Gehlenborg[1*]

[1]Harvard Medical School, Boston, MA, United States of America

*Corresponding Author: nils@hms.harvard.edu

**Abstract**

The number of data sets in biomedical repositories has grown rapidly over the past decade, providing scientists in fields like genomics and other areas of high-throughput biology with tremendous opportunities to re-use data. Scientists are able to test hypotheses computationally instead of generating their own data, to complement their own data sets with data generated by others, and to conduct meta analyses across many data sets. In order to effectively exploit existing data, 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. In order to address these challenges, we have developed SATORI—an ontology-guided visual exploration system—that combines a powerful metadata search with a tree map 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 browsing of the repository. The requirements for SATORI were derived in semi-structured interviews with biomedical data scientists. We demonstrate its utility by describing several usage scenarios using a stem cell data repository, discoveries we made in the process of developing them, and an evaluation of SATORI with domain experts. We have integrated an open-source, web-based implementation of SATORI in the data repository of the Refinery Platform for biomedical data analysis and visualization (`http://refinery-platform.org`).

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# 1 Introduction

Molecular biology is a scientific domain that is rapidly accumulating data in hundreds of public databases and repositories [33]. In particular, the field of genomics has seen a rapid increase in data generation in recent years. This is primarily driven by the wide availability and falling costs for technologies that enable genome-wide measurements of biological samples in a high-throughput fashion. The implementation of data release policies stipulated by journals and funding agencies has resulted in the availability of tens of thousands of data sets.

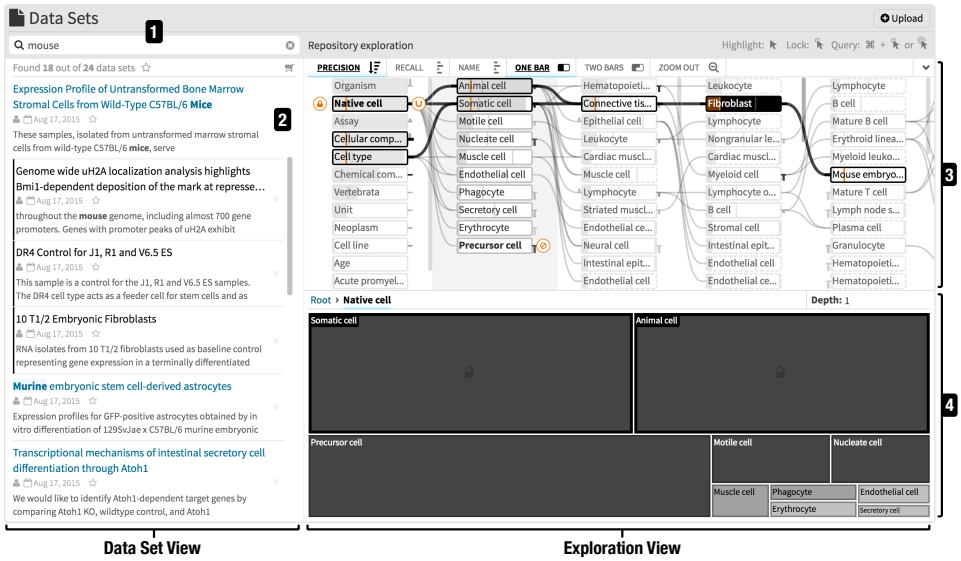


**Figure 1:** *SATORI illustrating a query for native cells excluding precursor cells combined with a synonym keyword search for mouse.The data set view includes the search interface (1) and the list of retrieved data sets (2). The exploration view is composed of the precision-recall plot (3) and the ratio plot (4), which show the abundance and other properties of annotation terms among the retrieved data sets. The black highlighting indicates the selection of fibroblast, its parent and child terms and the associated data sets.*

While all public data repositories are designed for data sharing, i.e. storage and dissemination of data sets, their usage and content vary. On the one end of the spectrum, there are *general purpose repositories* that hold data from many different studies. Their content is not collected for any particular project or purpose beyond data sharing. Typically such repositories exist for different biological data types, such as DNA sequences [44], gene expression data [5, 21], metabolomics data [13]

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

or proteomics data [46]. At the other end of the spectrum, there are *specialized data repositories* created for particular large-scale studies such as ENCODE [39] or the Roadmap Epigenomics Project [34]. These repositories are typically smaller and often contain multiple data types and use a limited set of metadata annotations.

As public data release is motivated by calls for transparency and reproducibility, not only experimental data is being released, but the data is also annotated with metadata describing the studies and properties of the analyzed samples and of the assays themselves. In this context and for the purpose of this manuscript, a data set is a collection of data files, including both raw and processed data, along with the corresponding metadata (see Figure 2). Metadata comprises the description of the overall study such as the aims, the experimental protocols used, and publications associated with the data, as well as the attributes of individual files, which typically correspond to the results of a particular biological assay such as sequencing, mass spectrometry, or gene expression analysis on a biological sample.

Standardized file formats and data structures have been developed to describe and exchange data set descriptions and metadata, e.g. MAGE-Tab and ISA-Tab [31, 36]. Additionally, the content of many repositories is maintained by expert curators, who ensure that at least the minimum information for reproducibility of a study is provided by submitters [7, 43] and that the metadata is not ambiguous [4]. In addition to checking free text used to describe study aims or protocols, curators also map metadata annotations to ontology terms or controlled vocabularies (see Figure 2) to facilitate semantic organization and retrieval of data sets.

Published data sets also offer tremendous opportunities for re-use of data for other purposes than for which they were originally generated. For example, in some cases individual published data sets can be used to test a hypothesis instead of generating new data. Alternatively, data from previous studies can be employed as corroborating evidence for observations made in an experiment. Meta studies that include data from dozens or hundreds of published data sets are another frequent use case for the re-purposing of previously generated data. For example Lukk et al. studied patterns of gene expression in human tissues based on hundreds of public gene expression data sets [24] and a similar study was conducted for mouse tissues by Zheng-Bradley et al. [49]. Other groups have studied connections between different diseases using publicly available data sets [8, 9, 42].

In order re-use published data sets from efficiently and to plan and execute aforementioned meta studies, it is necessary that scientists can gain an overview of the data that is contained in a data repository and can identify data sets that are related to a given topic of interest. If the goal is to conduct a study on common molecular patterns across diseases, it is crucial that a given data repository contains both data from a sufficiently large number of diseases as well as a sufficiently large number of data sets per disease.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

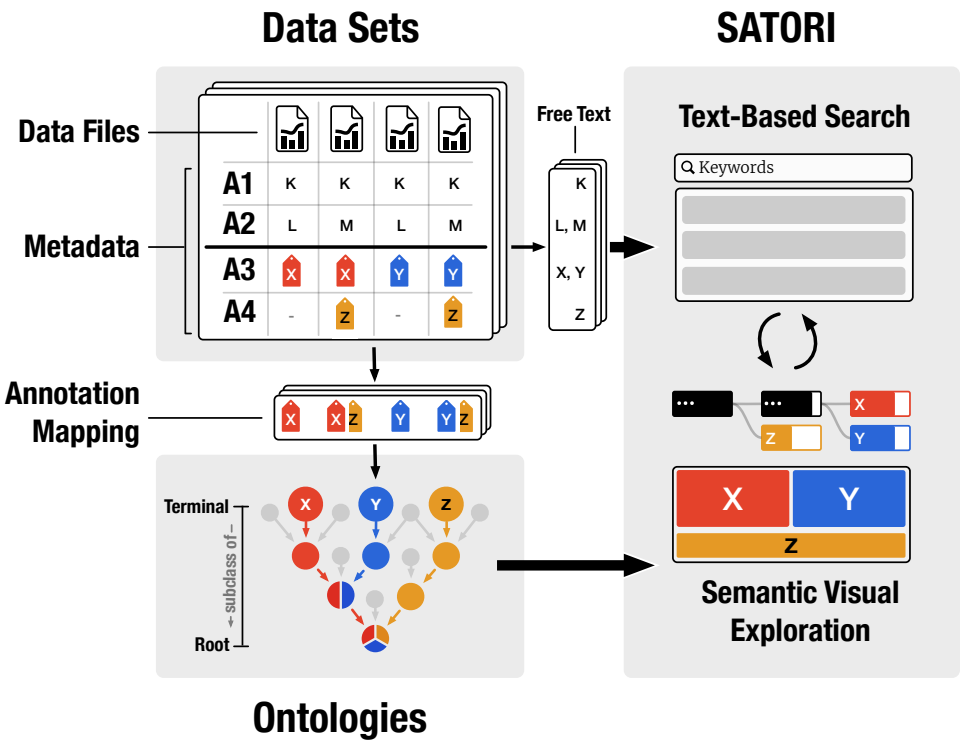


**Figure 2:** *Metadata attributes, such as creation date, technology, organism or disease, can contain free text and additionally ontologically annotated attribute values, which describe the set of data files contained in a data set. The free text is extracted, bundled into one text document and indexed for text-based search. The ontological annotations are extracted and linked to the ontology terms. Visual semantic exploration is performed on these ontology terms, e.g. X, Y, and Z, as well as all associated parent terms.*

There are many biomedical data repositories that provide some basic tools for data exploration. Most provide a comprehensive text-based search interface while a few also provide other means of exploring. For example, the two major data repositories for gene expression data are Gene Expression Omnibus (GEO) [5] and ArrayExpress [21], each containing over 60,000 submitted data sets as of March 2016. Both provide only standard text-based faceted searching. GEO has an indented list facet view for taxonomy groups and provides a list-based repository browser of high-level features such as sample types or organisms. ArrayExpress has one combined interface for exploring search results and browsing that features list-based filter options. Both of these interfaces are insufficient to fully address

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the needs of scientists for the re-use of published data.

Visual analytics approaches for information foraging have been shown to improve the efficiency of analysts in search and exploration [28] and here we present a novel visual analytics tool called SATORI (short for **S**emantic **A**nno**T**ations and **O**ntological **R**elations **I**nterface) that combines search and exploration. SATORI implements the foraging loop of the sensemaking process [29]. Our work makes two major contributions:

1. The description of three typical user roles for biomedical data repositories—data analyst, group leader, and curator—for which we derived needs and tasks in semi-structured interviews.

2. The creation of a unified visual exploration system that combines keyword-based search (free text and metadata fields) and metadata annotation term browsing with ontology guidance.

Furthermore, we are also providing an open-source implementation of SATORI within the Refinery Platform, which combines a data repository, analysis pipelines, and visual exploration tools. Using this implementation and a collection of almost 200 publicly available genomics and epigenomics data sets, we validated our approach in a field study with 6 domain experts.

# 2 Goals, Needs and Tasks

## 2.1 User Roles

Both authors have previously worked on data repositories and are familiar with biomedical informatics research. Therefore we are aware that data repositories are used by different types of users or that users can take different roles, which we characterized prior to undertaking our task analysis. We identified the following three primary user roles in the context of exploration of biomedical data repositories:

**R1** data analyst

**R2** group leader

**R3** data curator

These user roles are not mutually exclusive and hence a single person can, for example, act as either data analyst or a group leader at the same time.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

## 2.2 Needs

The primary concern of *data analysts* is turning experimental data into information and subsequently transform it into knowledge by answering questions of interest. Given biomedical data, a data analyst may be searching for data sets that are most relevant to a given biological problem or question. Precise description of the experimental characteristics is most important to assess relevance. While the characteristics that matter most can vary greatly depending on the project, the goals for finding relevant data sets are often similar such as finding data to test the validity of a hypothesis, complementing in-house generated data to improve confidence, compare quality between different data sets, check the robustness of an algorithm, or to broaden the scope a study. In order to achieve these goals the data analyst needs to:

**N1** find data sets that match certain experimental characteristics.

**N2** find data sets that are similar (or dissimilar) to given data sets.

The *group leader* is somebody who leads a group of data analysts and is mainly focused on finding collections of data sets. This could be a professor trying to plan a new study or a grant proposal. Here the foremost goal is to ensure that the group has access to data that will allow it to address new challenges in their scientific area. To achieve this, group leaders need to be able to find out whether a repository contains data sets matching certain experimental characteristics of interest, to get an overview of the current state of the repository, and also to discover trends in data generation and availability. Therefore, the primary need of a group leader is to:

**N3** get an overview of the distribution of the experimental characteristics across a collection of data sets.

Data curators are responsible for the quality of the metadata used to describe data sets, including descriptive free text as well as ontological annotations. Their needs are revolving around the current state of curation and how its quality can be improved to increase the expressiveness and findability [25] of data sets. Data curators are not concerned about finding specific data sets for analysis or discovering trends in data availability but instead care about the overall distribution and usage of annotation terms. Hence, data curators primarily need to:

**N4** get an overview of the annotation term hierarchy and term usage.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

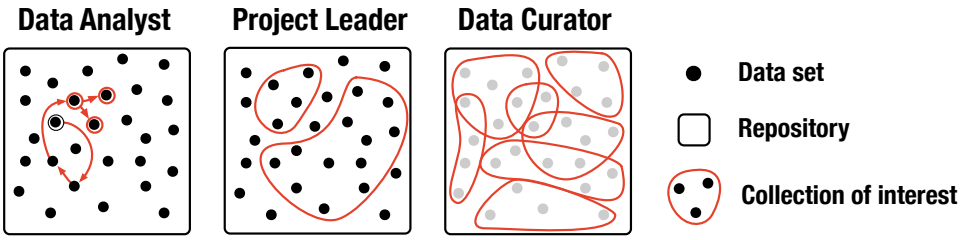


**Figure 3:** *Exploration behavior of different user roles. Data analysts aim at locating specific data sets. Group leaders focus on collections of data sets and the big picture. Data curators are primarily interested in the overall annotation term hierarchy instead of data sets.*

## 2.3 Requirement Analysis and Tasks

To develop a better understanding of the data discovery behavior in this domain, we performed a field study with eight PhD-level scientists and one graduate student from the biomedical informatics who had varying degrees of biomedical and computational expertise. Via semi-structured interviews we gained qualitative insights in their exploration behavior when searching for biomedical data. The results of these interviews guided our requirements analysis and design of the data set summary panel described in section 5.3. We identified a set of 10 tasks for the three user roles that a repository exploration system needs to support.

In order to meet the needs described above, some degree of understanding of the repository and subsets of it is required. Subsets can either represent text-based search results or ontology term-based queries. Understanding the composition of characteristics, i.e., annotations, is crucial for planning what to explore next. Hence, the requirements are separated into those that a relevant for interpreting the term composition of subsets and those that are concerned with finding subsets of interest in the the data repository.

The following five tasks are related to understanding what is contained in a collection of data sets:

**T1** Determine annotation terms of a data set.

**T2** Determine abundance of annotation terms of a group of data sets.

**T3** Determine abundance of sets of annotation terms among a group of data sets.

**T4** Determine annotation term containment relationships.

**T5** Summarize and view the metadata of a data set.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

The notion of relevant data sets (N1), i.e., data sets that significantly match a desired experimental characteristic, .e.g., such as "cancer", can be achieved through the illustration of annotation terms (T1) and by viewing summaries of data set descriptions (T5). Showing the relationship between annotations terms (T4) can facilitate finding of related data sets (N2). The abundance of single ontology terms (T2) and sets of ontology terms (T3) aids the understanding of search results and can highlight trends (N3, N4). Previewing certain details of a data set (T5) can further increase or decrease relevance and help to find the desired data (N1, N2).

The following four tasks are related to the process of exploring data collections:

**T6** Search for data sets.

**T7** Query data repository by annotation term.

**T8** Filter down a group of data sets according some annotation.

**T9** Loosen annotation constraints.

Being able to search (T6) by keywords and query (T7) by annotation terms for data sets is crucial in finding specific data sets as well as groups of data sets (N1, N2, N3, N4). Drilling down into search results by filtering according to some annotation terms (T8) and drilling up by loosening constraints (T9) supports exploration through an enriched search. Finally, ranking annotations according to their abundance and size enhances both: understanding of the nature of data sets (N3, N4) by highlighting most abundant or most scarce terms, and facilitates exploration by providing a notion of information scent [30]:

**T10** Rank annotations.

# 3 Related Work

Ontology-guided exploration of biomedical data repositories intersects with a number of different research areas. To provide a comprehensive overview of related work, we review two main areas: search visualization as well as tree and graph visualization methods.

## 3.1 Methods for Search Visualization

Exploration of data repositories or large document collections have similar requirements and goals as search visualization, since search results represent an arbitrary subset of the document corpus. Additionally, search visualization incorporates the notion of relevance for search-related retrieved data sets.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Over the last two decades, various different visualization methods have been developed to support search. The tools can be divided into those that attempt to visualize each result separately and those that try to provide an overview of the complete search result space. For example, TileBars [15] and its successors Insyder [32], and HotMap [18] visualize the approximated location of query term matches within each retrieved document and thus provide a visual notion of relevance. Others illustrate the relative similarity of search results by depicting each retrieved document as a glyph or simple visual mark in a 2D or 3D space, where the spatial location is mostly determined via dimensionality reduction. Similar documents should cluster together and form fuzzy groups or categories. Examples for glyph-based visualization techniques that operate on search results are InfoSky [2] and xFind's VisIslands [1]. InfoSky incorporates hierarchical classification of the documents and displays them in circular weighted Voronoi tree maps.

On the other hand, some visualization methods provide an abstract summary of the set of all retrieved documents. The RelationBrowser++ [47] visualizes the overall and search related abundance of categories using superimposed bar charts on category labels. The search engine Grokker categorized search results hierarchically and provided a top-down filter mechanism via a circular tree map of topics and subtopics. Note that Grokker was shut down when Groxis ceased operations in 2009. The Internet Archive provides a copy of Grokker's tour, which contains screenshots and a brief explanation of their visualization tool (`https://web.archive.org/web/20090509164021/http://www.groxis.com/service/grokker/grokker_tour.html`). ResultMaps [10] groups search results according to a hierarchical classification and uses the tree map visualization technique to convey the hierarchy. Hearst [14] provides a comprehensive overview of the efforts in visualizing search results.

Apart from that, a number of projects have studied possibilities to visually summarize the corpus of data repositories and enable exploration. The following examples focus on visual exploratory tools that utilize metadata or descriptive vocabulary, i.e. tools that visualize categorized or set-typed data. InfoSky [2] that has been described above can also be used to explore the data collection as a whole. Hiérarchie [40] is a tool for visualizing hierarchical topic models using sunburst charts for exploration of text documents. In a similar fashion, PhenoBlocks [12] uses the SunBurst idiom to represent the hierarchical structure of the Human Phenotype Ontology [35].

Even though SATORI only visualizes a very limited subset of the complete ontologies, it is worth mentioning that many efforts went into visualizing ontologies in their entirety. Katifori et al. [20] provide an extensive overview of different visualization techniques. Most methods focus on a representation for certain details, e.g. indented lists can show direct subclass and superclass relationships. On the

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

other hand, VOWL [23] is a specification for visualizing the complete structure of an ontology using the node-link idiom and predefined visual marks for the different aspects of an OWL ontology.

### 3.2 Methods for Tree- and Graph Visualization

Other work that indirectly relates to SATORI is more focused on visualization techniques for graph, tree, or hierarchical containment data. The variety of tree visualizations alone is huge. Hans-Jörg Schulz maintains an extensive collection of numerous different visualization methods for tree data [37]. Tree maps [19] are one of the most space efficient ways to visualize hierarchical data and have been studied extensively. A major disadvantage of tree maps is that they do not communicate the tree topology as well as a node-link diagrams. Elastic Hierarchies [48] have been developed to combine the strength of node-link diagrams and tree maps. GrouseFlocks [3] is another attempt to combine the node-link idiom with circular tree maps. Jigsaw's [41] list view illustrates relatedness of different items in lists via color coding and linking of items that are related to the one queried item. Parallel tag clouds [11] arrange feature words of different text corpora in parallel lists and highlight identical feature words via indicated links that get fully drawn once the user interacts with the visualization.

## 4 Data

Biomedical data sets are collections of primary data often consisting of various different file types. These files can be bundled and annotated with metadata to describe a complete study. Common formats are Microarray Gene Expression - Tabular format (MAGE-TAB) [31] or Investigation-Study-Assay Tabular format (ISA-Tab) [36]. Data repositories are web-based services that allow users to deposit annotated data sets for storage and retrieval by others.

Since free text metadata files allow the data generator to describe the content in any way, it is often unavoidable to introduce ambiguities. Controlled vocabularies or ontologies provide a means of resolving this issue by allowing the data generator to describe the content with a predefined set of terms.

### 4.1 Data Abstraction

The details of data types and structures of biomedical data can vary greatly depending on the research field and application but the fundamental components for ontology-guided exploration stay the same. Since the goal of this work is to find data sets rather than single data files, from an exploration point of view a data set is

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

regarded as an atomic unit with multiple characteristics, meaning that this project only focuses on inter- and not intra-data set exploration. A subset of the characteristics are linked to ontology terms. Given the transitive nature of parent-to-child term relationships in ontologies, characteristics that are linked to ontology terms are indirectly associated with the parent terms of that term as well. For example, the term *podocyte* is a child term of *epithelial cell*, which in turn is a subclass of *cell*. Hence, *podocyte* is also a child term of *cell* and all data sets annotated with *podocyte* are automatically annotated with *cell* as well.

## 4.2 Data Processing

Ontologies can be seen as directed, and in most cases acyclic, graphs where terms are represented by nodes and relationships are indicated via edges between two nodes. In the context of repository exploration, the most important property is the number of data sets that are associated with a term. Given a graph $G = (V, E)$ with $V$ representing the set of vertices and $E$ representing the set of edges, we denote the number of times a term $t$ has been used to annotate a data set as the *size* of the term. Terms can also be regarded as sets of data sets; given a term $t$, its set representation is denoted by $S_t$. Since ontologies describe *OWL:Thing* . By having a unique root node, we expose an indirect order on the node set. The length of the shortest path of a term $t$ to the root term is defined as the distance of $t$.

Most ontologies used for annotation in the biomedical domain each describe a different domain extensively but the number of terms that are used for annotation can be very limited compared to the overall number of terms. For example, the Stem Cell Commons data collection that is used across section 7 uses only 142 out of 1,269,955 terms. Since the goal of SATORI is to provide means for finding data sets and understanding the composition of data collections, rather than visualizing the ontologies themselves, terms that have not been used for annotations should not be shown. It should be noted that indirect annotation terms, i.e. parent terms of a direct term, should not be removed. But even the number of indirectly used terms might be high given the deep hierarchical structure of some ontologies. Each parent term of a term should account for a larger collection of data sets to provide an efficient exploration experience. Semantically, higher level term that describe the same collection of data sets could safely be hidden. For example, if a repository contains 10 data sets in total and all are related to *human* then the term *Mammalia* will describe the same 10 data sets, hence the mutual information of all parent terms of *human* related to other attributes (e.g. disease) is zero. The annotation term hierarchy should be a strict containment set hierarchy. Given three terms $A$, $B$ and $C$ where $A$ is a subclass of $B$ and $B$ is a subclass of $C$. The set representations $S_A$, $S_B$ and $S_C$ of the terms must fulfill:

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

$$S_A \subset S_B \subset S_C \tag{1}$$

This leads to pruning of terms whose size is zero as illustrated in Figure 4.1. For example, the Stem Cell Commons data collection (see section 7) includes data sets with files sampled from three different species: *human*, *mouse*, and *zebrafish*, which have all been annotated with the NCBITAXON ontology. Pruning the subgraph starting from the last common ancestor (*euteleostomi*) according to equation 1 results in the removal of 37 terms (Figure S2).

In addition, the tree map visualization method conveys the hierarchical order by containment; hence, data to be visualized needs to be provided in form of a tree rather than a graph. Terms with multiple parent terms and non-terminal terms with a size greater than zero are duplicated as illustrated in Figure 4.2. Although the precision-recall plot as described in Section 5.2.2 uses the visual metaphor of node-link diagrams, the hierarchy is illustrated by placing nodes from left (the root term) to the right (terminal terms). Depending on the complexity of the graph, it is possible that a node could be placed in multiple different columns, as there might be different path to the very root. To avoid visual clutter links only go in one direction: from the parent (left) to child (right). Thus, nodes with multiple parents whose distances to their parents are not equal, have to be duplicated (see Figure 4.3).

# 5 Design

SATORI is composed of three main interlinked views: the data set view, the exploration view, and a data set summary view. The first two components are visible by default and shown in Figure 1. The data set summary view is only shown on demand (Figure S3).

The data set view contains a list of data sets, which can contain all data sets of the repository, the retrieved data sets of a keyword search, or the result of a term query. The view includes the search interface consisting of a simple keyword query input at the top. The data set view also contains basic controls to filter and sort the list as well as a data cart, which allows the user to temporarily save data sets for later investigation.

The exploration view contains the ratio and precision-recall plot, which are the two main visualizations for understanding and exploring the data repository based on the annotation terms. The ratio plot uses the tree map technique to illustrate the term frequency ratios while the precision-recall plot uses a horizontal tree-like node-link diagram. Both plots display the same data but represent the attributes differently to compensate for each others limitations. While the tree map provides

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

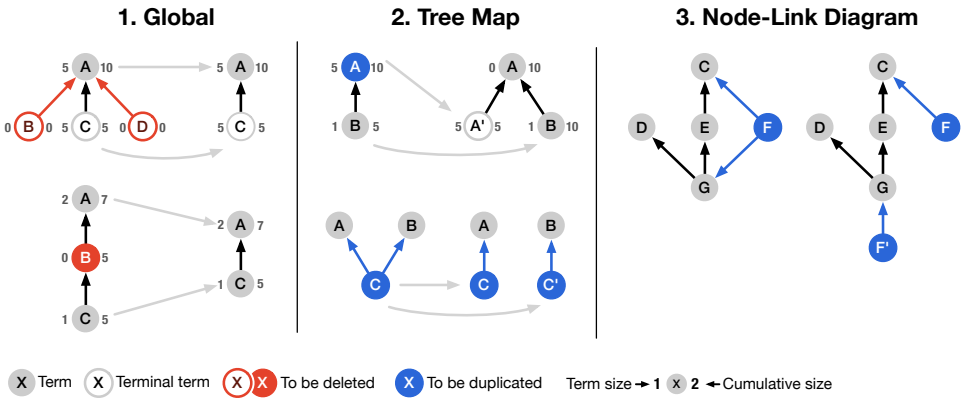


**Figure 4:** *Graph manipulations. (1) Leaves and inner nodes of size zero are deleted. The cumulative size includes the sum of the size of all child nodes. (2) Node duplication for the tree map visualization. Inner nodes with a size greater than zero need to be duplicated as child nodes to themselves. Also, nodes with multiple parents are duplicated for each parent to provide a unique path to the root. (3) The list graph visualization only requires nodes to be duplicated when the distances of their parents to the root are not the same. Therefore, node F is duplicated because the distance of node C and G is not equal. On the other hand, node G is not duplicated because the distance of nodes D and E is the same.*

a space-efficient overview of higher-level terms, the node-link diagram represents the actual relationships between terms across multiple levels.

In the ratio plot, a term is represented by a rectangle. The area of the rectangle visualizes the size of the term relative to its sibling terms. The color indicates the distance to farthest child term, i.e. the subtree depth. The farther away a child term is, the darker the rectangle. The precision-recall plot represents terms as nodes and visually links parent and child terms according to subclass relationships defined by the ontologies. Additionally, the precision-recall plot shows the precision and recall for each term given the currently retrieved data sets. Precision is defined as the number of data sets annotated with a term divided by the total number of retrieved data sets. Recall is defined as the number of retrieved data sets that have been annotated with the term of interest divided by the total number of data sets annotated with that term across the whole repository. In this context, precision is useful to understand how frequently a term is used for annotation in the retrieved set of data sets while recall provides a notion of information scent [30] by indicating if there are more data sets with that specific annotation term in the repository. Both plots also provide means of querying the data repository for annotation terms.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Finally, the data set summary view provides an overview of what a data set is about given and incorporates non-ontological metadata as well.

## 5.1 Data Set View

The data set view consists of the text-based search interface, data set list, and data cart. The search interface has been kept at a minimum in order to provide an easy-to-use interface [26]. A data set is represented by a surrogate holding the title of the data set, ownership and sharing information, and an indicator whether the data set is currently saved in the *data cart*. Additionally, search results feature a *keyword in context* (KWIC) snippet to show the context of the matched keywords (Figure 5). Browsing the whole repository, search results, or term-based query results are handled through the same interface to keep the learning process at a minimum. A search is initiated as soon as the user enters more than two characters and the results are displayed progressively as the user types. The minimal interface also helps to avoid errors and is in line with most other search interfaces. Providing a powerful search is crucial to address N1, N2 and the related task T6.

A click on the button with a double arrow icon (»), which is located right to the title, will open the data set summary view. Being able to quickly get a summary of the meta information regarding a data set is crucial as it helps the user evaluate the relevance of a retrieved data set (T5). SATORI follows the *Visual Information-Seeking Mantra* [38] and provides different levels of overviews: starting with the data set surrogate (Figure 5 and 6), a metadata summary (see Figure S3) and complete data set page (not shown here as this is part of the host application, i.e. the Refinery Platform in our case).

The data cart is contained within the data set view and enables the user to temporarily collect data sets of interest during the exploration process. The rationale behind the data cart is to reduce the cognitive load during search and provide a way to compare results from different searches or term queries. Data sets can be

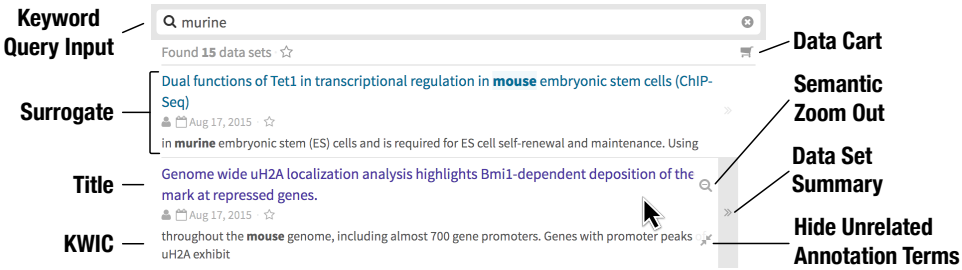


**Figure 5:** *Search interface with data set surrogates.*

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

added or removed to the data cart through a click on the outlined star icon in the data set surrogate (☆), which will subsequently get filled (★) and colored in yellow. The star icon right below the search input acts as an indicator and action button at the same time. A click will add all currently retrieved data sets to the data cart, e.g. in Figure 6 a click would add all 15 data sets to the cart. At the same time the icon indicates whether none (☆), some (⯪), or all (★) currently retrieved data sets have been added to the data cart. The data cart is explicitly designed to look very similarly to the data set list to indicate that their features are identical. This keeps the learning process at a minimum.

## 5.2 Exploration View

The data set view alone is most useful for navigational (e.g. to access a known or the most recently imported data set) and for some transactional (e.g. to find the owner of a known data set) search approaches. It works well in finding data that has been explicitly described in the free text metadata (Figure 2). On the other hand, when the exact context of a data set is unknown, a keyword-based search often fails and it provides no overview of the distribution of data sets with certain attributes across the repository. Ontologies provide a rich context given the semantic descriptions of attributes and hence enable a more targeted and therefore more efficient exploration. In the spirit of Pirolli and Card's [29] sensemaking process, SATORI aims at an improved information foraging process by enriching the search process with attribute-based exploration that visualizes the context of data sets and provides means of semantic top-down exploration, which is a common approach for exploring unknown data or for analyzing collections of data sets as proposed by Patterson et al. [27].

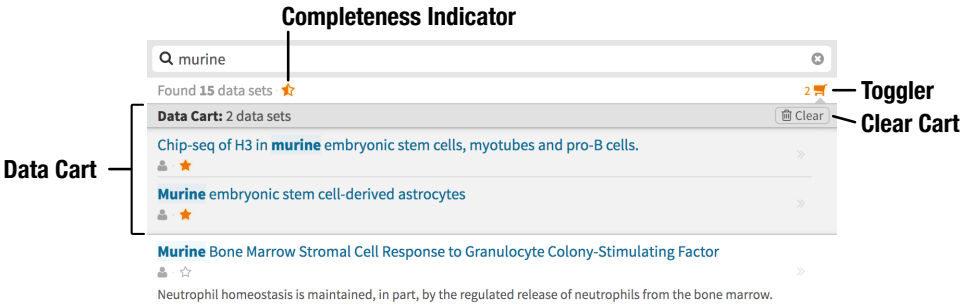


**Figure 6:** *Data cart for temporarily collecting data sets during the exploration process.*

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

#### 5.2.1 Ratio Plot

The ratio plot uses the tree map technique to visualize the size of the annotation term. The main advantage the tree map technique is that the currently selected tree level is always drawn within its container without any overflow. This provides an immediate overview. Other visualization techniques that are often used for deep hierarchies, e.g. indented lists or node-link diagrams, require some user interactions to see hidden parts, thus requiring the user to remember and recall hidden parts in order to reason about them. Three major disadvantage of tree maps are that it is hard to perceive the hierarchical structure [45], the rectangle areas are not relative to the root and engender ambiguities (Figure 7), and the area encoding is relatively imprecise compared to other encodings such as scaled length [16]. To compensate for these disadvantages we developed the precision-recall plot to complement the ratio plot.

#### 5.2.2 Precision-Recall Plot

The precision-recall plot uses the node-link technique to provide a strong visual notion of connectedness between related terms. Two terms are visually linked when they are related by the ontological *is superclass of* relationship. The directionality follows the reading direction of Latin languages, i.e. from left to right. For example, in Figure 7 it would be read that *A is a superclass of B*. To avoid visual clutter, only visible nodes are linked. To indicate the number and position of hidden links, a bar is displayed left or right of a term (▪) for incoming and outgoing links to nodes outside the visible area. The height of the bar indicates how much one needs to scroll the neighboring column to get to the linked nodes. The color indicates the amount of hidden links; the darker the gray, the more links are currently hidden.

Terms are ordered in columns by their distance to the root node and aligned to the top in order to increase the overall space efficiency (Figure S1). Each column is individually scrollable and the horizontal layout has been chosen over the vertical layout for a number of reasons: the flow of a path follows the reading direction, it is easier to compare *precision* and *recall* because the bars are aligned, and columns of terms (i.e. sibling terms) can be individually sorted and scrolled in a familiar fashion (e.g., vertical scrolling and sorting is a common feature in typical file explorers). A vertical layout would potentially provide a more focused view on siblings across different branches but that is already well supported by the depth control of the recall plot.

By default, the superimposed bar displays precision (e.g. ▬▭ precision is 50%) and the superimposed vertical line indicates recall (e.g. ▭▭ recall is 75%). The rationale behind this default behavior is that the user will in most cases start ex-

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ploring the complete repository first, hence recall is always equal to 1 and hence not informative. The top bar contains controls that allow the user to sort the nodes (ascending or descending) according to *precision*, *recall*, and by *name*. The *one bar* and *two bar* controls allow to switch between one superimposed bar and a vertical line indicator (▬) and two superimposed bars (▬) for *precision* and *recall*. It is also possible to sort columns individually. The *zoom out* button allows to decrease the size of the complete plot to inspect the overall structure of the graph.

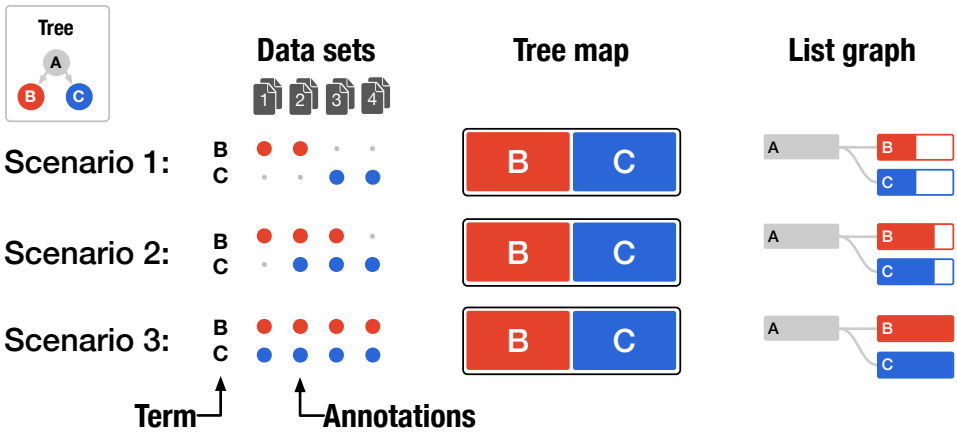


**Figure 7:** *Since the area of a the rectangle of a treemap reflects the ratio among all other rectangles, it is not always possible to conclude how the area compares to the root, especially when terms represent intersecting sets of data sets. In this example, the treemap looks identical in all three scenarios even though the actual sizes of the terms differ. The superimposed bar charts of the precision-recall plot avoid this problem.*

## 5.3 Data Set Summary

The data set summary view is crucial in supporting the reading and information exporting step of the information foraging loop [29] and to address T5. The layout has been designed to reflect the priorities for data set properties that we derived from the initial semi-structured interviews with domain experts (Table S2 and S3).

## 5.4 Interactions and Querying

All components of SATORI are highly interlinked to provide an integrative exploration experience and to visualize the context of retrieved data sets given the ontological annotations.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

### 5.4.1 Identify Data Set Associated Annotations

When hovering over a data set in the data set list, all associated annotation terms are highlighted in the ratio and precision-recall plot by changing the hue of the term to orange. Since the precision-recall plot shows nodes across different levels, direct and indirect annotation terms are handled differently. Direct annotation terms are those that have been used to directly annotate an attribute value of the data set and are filled in orange. Indirect annotation terms include all parent terms, e.g. *motile cell* and *native cell*. Highlighting the data set terms addresses T1.

While the tree map technique used for the ratio plot always displays all terms on a certain branch or level, the size of the node-link diagram used as the precision-recall plot might exceed the size of the visible area, hence some parts can be occluded. In order to focus on the terms of a hovered data set, the user can semantically zoom out, when inspecting a data set, via a click on the button with a magnifier (🔍) to such an extend that all annotation terms related to the hovered data set are visible. Furthermore it is possible to hide all unrelated annotation terms via a click on the button below the semantic zoom out (↖). Figure 5 shows an example with both buttons.

### 5.4.2 Discover Data Set By Annotations

To discover and explore the repository based annotated attribute values SATORI highlights associated data sets with a black border and slight shift to the right (Figure 8) when hovering a term in either of the two plots in the exploration view. To make the highlighting persistent, it can be locked via a click on a rectangle in the ratio plot or by clicking on the lock button of the context menu of the precision-recall plot (Figure 9).

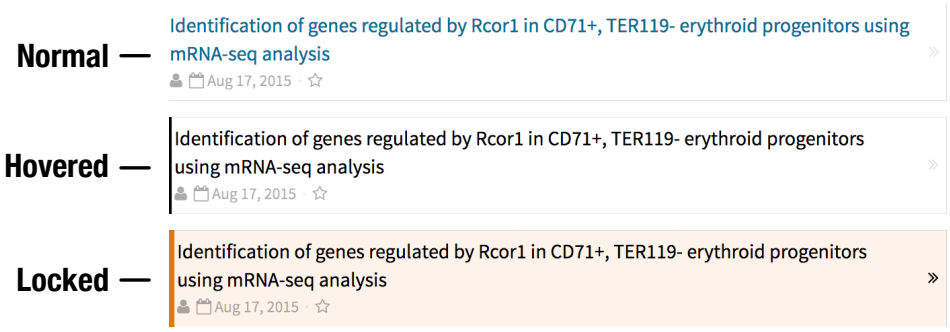


**Figure 8:** *The three possible states of a data set surrogate in the data set list view. Hovered and locked states can also be combined.*

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Both visualization plots support term-based querying to address T7, T8, and T9. A double click on a term in the ratio plot will zoom into the subtree and simultaneously restrict the data set to be associated with this subtree, hence the data set collection is queried for the clicked term (T7 and T8). At the same time, the precision-recall plot sets the clicked term as its new root. The same action can be triggered in the precision-recall plot via a click on the *Root* button in the term context menu. Loosening annotation constraints (T9) is can be achieved through a click on the ratio plots breadcrumb-like root path view (Figure 10 or by clicking again on the *Root* button in the term context menu.

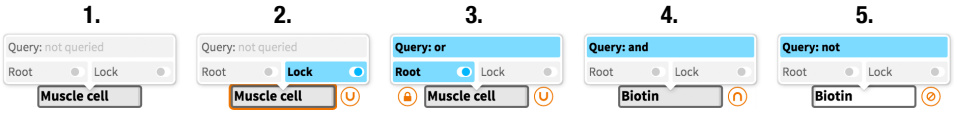


**Figure 9:** *The term context menu (1) of the precision-recall plot controls three interaction: visually locking term related (2), re-rooting the graph (3), querying the data collection by a term (3, 4, and 5).*

Additionally, The precision-recall plot supports more complex Boolean annotation term queries via the term context menu's query button. The query button features four query states: *none*, *or*, *and*, and *not* which the user can toggle through by clicking multiple times on it. A blue indicator bar that fills the button in a blue hue top-down indicates the time left before the query is triggered. Since a query is likely to influence the composition of the retrieved data sets, *precision* and *recall* are likely to change and subsequently the position of terms in the node link diagram as well. The throttling options gives a user the ability to quickly toggle through the query options with having to search the term every time a query was issued.

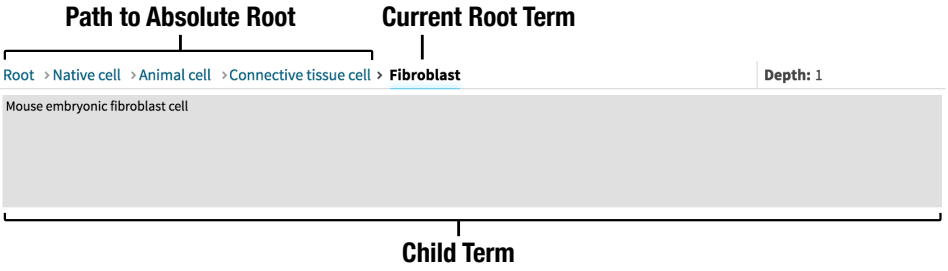


**Figure 10:** *The breadcrumb path to the absolute root term used in the ratio plot, which is useful for drilling up.*

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

# 6 Implementation and Scalability

SATORI is a web-based exploration system. The front-end is implemented in JavaScript using D3.js [6] and AngularJS. The information retrieval system is powered by Solr and ontologies are managed by a Neo4J graph database. While Solr manages and provides access to the metadata, Neo4J stores the complete ontological graph. A custom Java plug-in provides access to and retrieves the user specific sub graph for visualization. The Refinery Platform application manages the data set collection and controls the business logic between Solr and Neo4J. All parts of SATORI are open-source, publicly accessible at `https://github.com/parklab/refinery-platform` and `https://github.com/flekschas/d3-list-graph`, and continuously integrated via Travis-CI to ensure correctness and compatibility. Even though SATORI is implemented in the Refinery Platform (`http://refinery-platform.org`), it can be integrated in into other web-based data repositories as well.

Since biomedical data repositories can grow quickly, scalability is an important property of any repository exploration system. The performance of SATORI foremost depends on the total number of ontology terms that are used for annotation. The impact of the number of data sets or ontologies used is negligible since Solr and Neo4J are capable of handling millions of documents or nodes. Only the data of visible data sets is cached and our custom Neo4J plug-in fetches only the terms that are directly or indirectly used for annotation. Therefore the actual size of the global ontology graph does not affect retrieval performance. Currently, the precision-recall plotis the limiting factor as it displays the full annotation graph. We have tested the tool with up to 1000 annotation terms and while the performance decreases, the tool still remains usable.

# 7 Evaluation

To evaluate the utility of SATORI in understanding and exploring a biomedical data repository we asked six domain experts to work with SATORI on a real-world data collection. Moreover, the Supplementary Video provides a usage scenario based on the same data collection. We utilized the expert-curated data collection of the Stem Cell Commons [17] project, which is comprised of 199 ontologically annotated stem cell data sets. The data sets have been annotated with terms from 12 biomedical ontologies (see Table S1).

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

### 7.1 Field Study

A field study was conducted with five PhD-level scientists and one graduate student from the biomedical informatics domain. It consisted of a brief introduction to SATORI, a set of questions related to the exploration of the Stem Cell Commons data collection, and finally recording of anecdotal evidence about SATORI. We selected the participants to cover all identified user roles (Section 2.1). Three of the six participants were recurring participants from the initial interviews described in section 2.3. Since we assume that SATORI would mostly frequently be used by data analysts (R1), we chose to include four data analysts, one group leader (R2), and one data curator (R3).

The introduction briefly covered all important aspects of SATORI to give the participants an overview of the current feature set. The goal of this field study was to evaluate the general utility of our approach for exploration rather than to perform a usability test, which is planned for future evaluations; hence, an introduction allowed us to skip a lengthy familiarization with the system. We prepared some kick-off tasks (Table S3) in order to provide initial guidance in exploration but we also asked the participants explore the data on their own and beyond those questions. The rationale behind prepared set of tasks to focus on questions relevant for the Stem Cell Commons data repository.

All of the participants agreed that SATORI gives them a better understanding of the content of the overall repository compared to a system with only text-based search. Two data analysts mentioned that the ontology-guided exploration interface is very useful for collecting data sets associated with higher-level attribute values, which are not mentioned in the data set description (e.g. neoplasm as compared to glioma). The group leader mentioned that SATORI significantly aids exploration of unknown big data collections. The data curator said that "it is really exciting to finally see and explore the data" (in regards to the ontological annotations) and that it will be a useful asset for future data curation.

The greatest drawback of SATORI is that everyone agreed that it is currently very hard to locate specific terms within either of the two plots. Everyone would like to be able to search for annotation terms. Also, all participants mentioned at the end of the session that SATORI require some training or comprehensive introduction. Finally, everyone agreed that many high-level annotation terms are too generic and not useful for exploring the data repository as they are associated with all data sets. For example, the three terms: *thing*, which is the parent of *experimental factor*, which in turn is the parent of *material entity*, do not provide any insights. Based on this feedback we defined a set of more meaningful terms in regards to exploration, which can be seen in Figure 1.

We also asked the participants if the data set summary panel helps them to get

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a better idea of the data set content to evaluate whether our design choices—driven by the initial semi-structured interview—need improvement. All participants said that the summary view contains all of the essential attributes for determining the relevance of a data set.

# 8 Discussion

The feedback collected in our field study (Section 7.1) demonstrates that SATORI addresses the needs of the three defined user roles (Section 2.1) and successfully supports users the tasks (Section 2.3) in exploring a collection of data sets.

Two major drawbacks became apparent during the evaluation of SATORI: locating annotation terms is time-consuming and it is hard to comprehend the current state of querying. In order to address both problems at the same time, we propose a unified query interface, which should display and handle possible actions that manipulate the collection of retrieved data sets. Being able to search for terms would solve the issue of locating specific terms. The second issue could be addressed by displaying all query operation in one place. Hence, the unified query interface should handle text-based free text search, annotation term search, annotation term query operations, and basic filtering.

A common challenge in visualizing ontology-driven set hierarchies is that the ontologies define complex polyhierarchies. A trade-off has to be made between complexity and usability. We will look into ways how to resolve duplicated nodes in the tree map without altering the position of terms too drastically. Regarding the node-link diagram used in the precision-recall plot we are investigating ways to draw links across multiple columns without introducing too much clutter.

Another issue that came up during the evaluation of SATORI is that the current system requires some initial guidance or training before the tool can be fully exploited. A guided introduction tutorial or small interactive snippets could help to overcome this challenge.

Finally, comparing different groups of annotation terms (T3) is currently possible only indirectly. To address this limitation we will be looking into how existing techniques for exploring set intersections such as UpSet [22] could be integrated into SATORI to enable richer comprehension of term-related set properties.

# 9 Conclusion

SATORI is web-based exploration system that combines powerful search with visual browsing to provide an integrated exploration experience. The visualizations

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

serve two purposes: supporting the information foraging loop [30] and pattern discovery of attribute distributions as well as ontology-guided querying of the data repository.

SATORI contributes to the field of visual analytics by unifying a powerful text-based search with two exploration approaches that put data sets into context and shed light in the repository-wide distribution of biological attributes. We have identified three distinct user roles and evaluated their search behavior.

During the development and evaluation of SATORI we realized that—apart from the design of the visualization and the implementation—the greatest challenge of any semantic exploration approach is that its utility significantly depends on the quality of data curation. Fortunately, SATORI makes it easy to evaluate the current state of curation and find areas that need improvements.

We also learned that due to the nature of the complexities of ontologies, ontology-guided exploration tools require initial learning and are currently most useful for expert users.

## 10 Future Work

In future work we are planning to extend SATORI in multiple ways to further strengthen the integration of classic search and semantic visual exploration.

First and foremost, we want to evaluate the generality of our exploration approach. The biomedical research domain is known to deal with highly structured data but the question is whether other fields that work with non-textual data, e.g. video or music, could also benefit from semantic visual exploration.

An application that captures the individual steps of the whole exploration process could address the currently limited interplay between the search and the two visualization-based browsing methods. Often users go back and forth, change search keywords, or browse in different directions depending the previous results. Having a way to look at previous exploration steps, link their results, and provide an overview of the explored space without having to leave the current view could facilitate the understanding of consequences of each step and furthermore point out undiscovered areas of the data repository. Shneiderman described this feature in his *Information Seeking Mantra* [38] as the *history*.

It would also be interesting to study in which order annotation terms is preferable. Currently high-level parent terms are visualized first and the direct annotations terms can be found on demand but it might be more useful to show the direct annotation terms first and hide the higher level terms until the user requests them.

Finally, integrating non-ontological structured metadata into SATORI could have a huge impact as probably most descriptive metadata used nowadays is not

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

ontologically annotated.

## 11 Acknowledgments

We would like to thank everyone who participated in our interviews and the evaluation study to help us design and validate SATORI. Furthermore, we would like to express our gratitude to the members of the Refinery Platform team who helped to integrate and deploy SATORI. This work was funded by the National Institutes of Health (R00 HG007583) and the Harvard Stem Cell Institute.

## References

[1] K. Andrews, C. Gutl, J. Moser, V. Sabol, and W. Lackner. Search result visualisation with xFIND. In *Second International Workshop on User Interfaces to Data Intensive Systems, 2001. UIDIS 2001. Proceedings*, pages 50–58, 2001.

[2] K. Andrews, W. Kienreich, V. Sabol, J. Becker, G. Droschl, F. Kappe, M. Granitzer, P. Auer, and K. Tochtermann. The InfoSky Visual Explorer: Exploiting Hierarchical Structure and Document Similarities. *Information Visualization*, 1(3-4):166–181, Dec. 2002.

[3] D. Archambault, T. Munzner, and D. Auber. GrouseFlocks: Steerable Exploration of Graph Hierarchy Space. *IEEE Transactions on Visualization and Computer Graphics*, 14(4):900–913, July 2008.

[4] C. A. Ball, A. Brazma, H. Causton, S. Chervitz, R. Edgar, P. Hingamp, J. C. Matese, H. Parkinson, J. Quackenbush, M. Ringwald, S.-A. Sansone, G. Sherlock, P. Spellman, C. Stoeckert, Y. Tateno, R. Taylor, J. White, and N. Winegarden. Submission of Microarray Data to Public Repositories. *PLOS Biol*, 2(9):e317, Aug. 2004.

[5] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva. NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research*, 41(D1):D991–D995, Jan. 2013.

[6] M. Bostock, V. Ogievetsky, and J. Heer. D3: Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[7] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. C. Causton, and others. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature genetics*, 29(4):365–371, 2001.

[8] J. Caldas, N. Gehlenborg, A. Faisal, A. Brazma, and S. Kaski. Probabilistic retrieval and visualization of biologically relevant microarray experiments. *Bioinformatics*, 25(12):i145–i153, June 2009.

[9] J. Caldas, N. Gehlenborg, E. Kettunen, A. Faisal, M. Rönty, A. G. Nicholson, S. Knuutila, A. Brazma, and S. Kaski. Data-driven information retrieval in heterogeneous collections of transcriptomics data links SIM2s to malignant pleural mesothelioma. *Bioinformatics (Oxford, England)*, 28(2):246–253, Jan. 2012.

[10] E. Clarkson, K. Desai, and J. Foley. ResultMaps: Visualization for Search Interfaces. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1057–1064, Nov. 2009.

[11] C. Collins, F. Viegas, and M. Wattenberg. Parallel Tag Clouds to explore and analyze faceted text corpora. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST '09)*, pages 91–98, 2009.

[12] M. Glueck, P. Hamilton, F. Chevalier, S. Breslav, A. Khan, D. Wigdor, and M. Brudno. PhenoBlocks: Phenotype Comparison Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):101–110, Jan. 2016.

[13] K. Haug, R. M. Salek, P. Conesa, J. Hastings, P. d. Matos, M. Rijnbeek, T. Mahendraker, M. Williams, S. Neumann, P. Rocca-Serra, E. Maguire, A. González-Beltrán, S.-A. Sansone, J. L. Griffin, and C. Steinbeck. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Research*, 41(D1):D781–D786, Jan. 2013.

[14] M. Hearst. *Search User Interfaces*. Cambridge University Press, Sept. 2009.

[15] M. A. Hearst. TileBars: Visualization of Term Distribution Information in Full Text Information Access. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '95)*, pages 59–66. ACM, 1995.

[16] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI*

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

*Conference on Human Factors in Computing Systems*, CHI '10, pages 203–212. ACM, 2010.

[17] S. Ho Sui, E. Merrill, N. Gehlenborg, P. Haseley, I. Sytchev, R. Park, P. Rocca-Serra, S. Corlosquet, A. Gonzalez-Beltran, E. Maguire, O. Hofmann, P. Park, S. Das, S.-A. Sansone, and W. Hide. The Stem Cell Commons: an exemplar for data integration in the biomedical domain driven by the ISA framework. *AMIA Joint Summits on Translational Science proceedings AMIA Summit on Translational Science*, 2013:70, 2013.

[18] O. Hoeber and Xue Dong Yang. The Visual Exploration ofWeb Search Results Using HotMap. In *Information Visualization 2006 (IV 2006)*, pages 157–165. IEEE, 2006.

[19] B. Johnson and B. Shneiderman. Tree-maps: a space-filling approach to the visualization of hierarchical information structures. In *Proceedings of the IEEE Conference on Visualization (Vis '91)*, pages 284–291, 1991.

[20] A. Katifori, C. Halatsis, G. Lepouras, C. Vassilakis, and E. Giannopoulou. Ontology visualization methods—a survey. *ACM Computing Surveys*, 39(4):10–es, Nov. 2007.

[21] N. Kolesnikov, E. Hastings, M. Keays, O. Melnichuk, Y. A. Tang, E. Williams, M. Dylag, N. Kurbatova, M. Brandizi, T. Burdett, K. Megy, E. Pilicheva, G. Rustici, A. Tikhonov, H. Parkinson, R. Petryszak, U. Sarkans, and A. Brazma. ArrayExpress update—simplifying data submissions. *Nucleic Acids Research*, page gku1057, Oct. 2014.

[22] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. UpSet: Visualization of Intersecting Sets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '14)*, 20(12):1983–1992, 2014.

[23] S. Lohmann, S. Negru, F. Haag, and T. Ertl. VOWL 2: User-Oriented Visualization of Ontologies. In K. Janowicz, S. Schlobach, P. Lambrix, and E. Hyvönen, editors, *Knowledge Engineering and Knowledge Management*, volume 8876, pages 266–281. Springer International Publishing, Cham, 2014.

[24] M. Lukk, M. Kapushesky, J. Nikkilä, H. Parkinson, A. Goncalves, W. Huber, E. Ukkonen, and A. Brazma. A global map of human gene expression. *Nature Biotechnology*, 28(4):322–324, Apr. 2010.

[25] P. Morville. *Ambient Findability: What We Find Changes Who We Become*. "O'Reilly Media, Inc.", Sept. 2005.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[26] J. Nielsen. *Usability Engineering*. Elsevier, Nov. 1994.

[27] E. S. Patterson, E. M. Roth, and D. D. Woods. Predicting Vulnerabilities in Computer-Supported Inferential Analysis under Data Overload. *Cognition, Technology & Work*, 3(4):224–237, Dec. 2001.

[28] P. Pirolli and S. Card. Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 51–58. ACM Press/Addison-Wesley Publishing Co., 1995. bibtex: pirolli_information_1995-1.

[29] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of International Conference on Intelligence Analysis*, 2005.

[30] P. Pirolli, S. K. Card, and M. M. Van Der Wege. The Effect of Information Scent on Searching Information: Visualizations of Large Tree Structures. In *Proceedings of the Working Conference on Advanced Visual Interfaces*, AVI '00, pages 161–172, New York, NY, USA, 2000. ACM.

[31] T. F. Rayner, P. Rocca-Serra, P. T. Spellman, H. C. Causton, A. Farne, E. Holloway, R. A. Irizarry, J. Liu, D. S. Maier, M. Miller, K. Petersen, J. Quackenbush, G. Sherlock, C. J. Stoeckert, J. White, P. L. Whetzel, F. Wymore, H. Parkinson, U. Sarkans, C. A. Ball, and A. Brazma. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics*, 7(1):489, Nov. 2006.

[32] H. Reiterer, G. Tullius, and T. M. Mann. INSYDER: a content-based visual-information-seeking system for the Web. *International Journal on Digital Libraries*, 5(1):25–41, Mar. 2005.

[33] D. J. Rigden, X. M. Fernández-Suárez, and M. Y. Galperin. The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection. *Nucleic Acids Research*, 44(D1):D1–D6, Jan. 2016.

[34] Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, Feb. 2015.

[35] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos. The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. *The American Journal of Human Genetics*, 83(5):610–615, Nov. 2008.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[36] P. Rocca-Serra, M. Brandizi, E. Maguire, N. Sklyar, C. Taylor, K. Begley, D. Field, S. Harris, W. Hide, O. Hofmann, S. Neumann, P. Sterk, W. Tong, and S.-A. Sansone. ISA Software Suite: Supporting Standards-Compliant Experimental Annotation and Enabling Curation at the Community Level. *Bioinformatics*, 26(18):2354–2356, 2010.

[37] H.-J. Schulz. Treevis.net: A Tree Visualization Reference. *IEEE Computer Graphics and Applications*, 31(6):11–15, 2011.

[38] B. Shneiderman. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the IEEE Symposium on Visual Languages (VL '96)*, pages 336–343, 1996.

[39] C. A. Sloan, E. T. Chan, J. M. Davidson, V. S. Malladi, J. S. Strattan, B. C. Hitz, I. Gabdank, A. K. Narayanan, M. Ho, B. T. Lee, L. D. Rowe, T. R. Dreszer, G. Roe, N. R. Podduturi, F. Tanaka, E. L. Hong, and J. M. Cherry. ENCODE data at the ENCODE portal. *Nucleic Acids Research*, 44(D1):D726–732, Jan. 2016.

[40] A. Smith, T. Hawes, and M. Myers. Hierarchie: Visualization for Hierarchical Topic Models. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 71–78. Association for Computational Linguistics, 2014.

[41] J. Stasko, C. Görg, and Z. Liu. Jigsaw: Supporting Investigative Analysis through Interactive Visualization. In *Proceedings of the IEEE Symposium on Visual Analytics in Science and Technology*, VAST '07, pages 131–138. IEEE, 2007.

[42] S. Suthram, J. T. Dudley, A. P. Chiang, R. Chen, T. J. Hastie, and A. J. Butte. Network-Based Elucidation of Human Disease Similarities Reveals Common Functional Modules Enriched for Pluripotent Drug Targets. *PLoS Computational Biology*, 6(2), Feb. 2010.

[43] C. F. Taylor et al. Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology*, 26(8):889–896, Aug. 2008.

[44] K. A. Tryka, L. Hao, A. Sturcke, Y. Jin, Z. Y. Wang, L. Ziyabari, M. Lee, N. Popova, N. Sharopova, M. Kimura, and M. Feolo. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Research*, 42(Database issue):D975–D979, Jan. 2014.

bioRxiv preprint doi: https://doi.org/10.1101/046755; this version posted April 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

[45] F. van Ham and J. J. van Wijk. Beamtrees: compact visualization of large hierarchies. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002*, pages 93– 100. IEEE, 2002.

[46] J. A. Vizcaíno, A. Csordas, N. del Toro, J. A. Dianes, J. Griss, I. Lavidas, G. Mayer, Y. Perez-Riverol, F. Reisinger, T. Ternent, and others. 2016 update of the PRIDE database and its related tools. *Nucleic acids research*, 44(D1):D447–D456, 2016.

[47] J. Zhang and G. Marchionini. Coupling browse and search in highly interactive user interfaces: a study of the relation browser++. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries*, page 384. ACM Press, 2004.

[48] S. Zhao, M. McGuffin, and M. Chignell. Elastic hierarchies: combining treemaps and node-link diagrams. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis '05)*, pages 57–64. IEEE Computer Society Press, 2005.

[49] X. Zheng-Bradley, J. Rung, H. Parkinson, and A. Brazma. Large scale comparison of global gene expression patterns in human and mouse. *Genome Biology*, 11(12):1–11, 2010.