Journal of the American Medical Informatics Association, 25(3), 2018, 337–344
doi: 10.1093/jamia/ocx134
Advance Access Publication Date: 30 November 2017
Research and Applications

Research and Applications

User needs analysis and usability assessment of DataMed – a biomedical data discovery index

Ram Dixit,[1] Deevakar Rogith,[1] Vidya Narayana,[1] Mandana Salimi,[1] Anupama Gururaj,[1] Lucila Ohno-Machado,[2] Hua Xu,[1] and Todd R Johnson[1]

[1]University of Texas Health Science Center at Houston, School of Biomedical Informatics, Houston, TX, USA and [2]University of California San Diego Health System, Department of Biomedical Informatics, La Jolla, CA, USA

Corresponding Author: Todd R Johnson, 7000 Fannin Street, Suite 800, Houston, TX 77030, USA. E-mail: Todd.R.Johnson@uth.tmc.edu. Phone: 713-500-3913

Received 30 May 2017; Revised 16 October 2017; Editorial Decision 19 October 2017; Accepted 27 October 2017

ABSTRACT

Objective: To present user needs and usability evaluations of DataMed, a Data Discovery Index (DDI) that allows searching for biomedical data from multiple sources.

Materials and Methods: We conducted 2 phases of user studies. Phase 1 was a user needs analysis conducted before the development of DataMed, consisting of interviews with researchers. Phase 2 involved iterative usability evaluations of DataMed prototypes. We analyzed data qualitatively to document researchers' information and user interface needs.

Results: Biomedical researchers' information needs in data discovery are complex, multidimensional, and shaped by their context, domain knowledge, and technical experience. User needs analyses validate the need for a DDI, while usability evaluations of DataMed show that even though aggregating metadata into a common search engine and applying traditional information retrieval tools are promising first steps, there remain challenges for DataMed due to incomplete metadata and the complexity of data discovery.

Discussion: Biomedical data poses distinct problems for search when compared to websites or publications. Making data available is not enough to facilitate biomedical data discovery: new retrieval techniques and user interfaces are necessary for dataset exploration. Consistent, complete, and high-quality metadata are vital to enable this process.

Conclusion: While available data and researchers' information needs are complex and heterogeneous, a successful DDI must meet those needs and fit into the processes of biomedical researchers. Research directions include formalizing researchers' information needs, standardizing overviews of data to facilitate relevance judgments, implementing user interfaces for concept-based searching, and developing evaluation methods for open-ended discovery systems such as DDIs.

Key words: data discovery, information retrieval, usability, user needs, metadata

INTRODUCTION

As the number, size, and public availability of biomedical datasets grow, so do the opportunities for new forms of research to advance biomedical knowledge.[1–3] However, the heterogeneous nature of biomedical data, the complexity of data-intensive research, and the lack of data discovery infrastructure pose significant challenges for researchers to take advantage of this opportunity.[4–6] Fragmented data environments, lack of data standards, and poor documentation are key issues that limit the direction and scope of data-driven research, often in the initial discovery phase.[1,5–7] Data must be better organized to facilitate the advancement of biomedical science.[1,5,8]

© The Author 2017. Published by Oxford University Press on behalf of the American Medical Informatics Association.
All rights reserved. For Permissions, please email: journals.permissions@oup.com

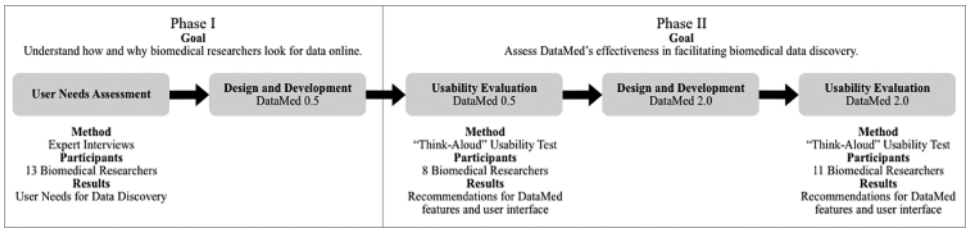

Figure 1. Diagram showing UCD process for DataMed. Phase 1 research was conducted prior to the development of DataMed; Phase 2 evaluations were conducted on versions 0.5 and 2.0 of DataMed.

In 2013, the National Institutes of Health Big Data to Knowledge (BD2K) initiative issued a call to assemble data from multiple sources into a discovery system termed a Data Discovery Index (DDI).[1] A DDI aims to accelerate data-intensive biomedical science by providing a mechanism for searching publicly available data.[1] A prototype DDI, DataMed, was launched in 2015 by the biomedical and healthCAre Data Discovery Index Ecosystem (bioCADDIE) project.[9] DataMed uses a common model (DatA Tag Suite [DATS], described elsewhere) to index metadata from biomedical data repositories, providing researchers with a PubMed-like search engine for discovering datasets relevant to their research interests.[10]

Aggregating datasets poses challenges in information retrieval and user interface design to enable the provision of meaningful search results to users.[11,12] While information seeking in literature indexes such as PubMed has been studied, biomedical researchers' information needs in data discovery are poorly understood and the conceptualization of data as an information resource in biomedicine is still emerging. Existing work has partially addressed the complexity and open-endedness of dataset search and evaluation, the diversity of research purposes and expertise, and the importance of context in understanding researchers' information needs.[4,5,13] Additionally, dataset retrieval differs from literature retrieval, because the heterogeneous sources and types of data and metadata make traditional indexing techniques insufficient.[12,15]

A successful DDI must meet researchers' information needs and fit into their research practices. User-centered design (UCD) ensures that technology meets users' needs through an iterative process of design and evaluation.[16] UCD aims to produce systems that are useful, usable, and effective. Usable systems complement users' knowledge, skills, and contexts, improving the effectiveness and satisfaction with which they accomplish their goals.[17] Usability evaluation in UCD consists of both quantitative and qualitative measures. However, existing quantitative and formal methods for evaluating information retrieval systems, such as precision and recall, do not adequately measure the subjective process of "discovering" data.[14–16] Qualitative methods such as user needs analysis, which involves understanding users' work goals and priorities, and usability testing, which consists of simulating representative tasks by representative users on system prototypes, are thus well suited for guiding the design and development of data discovery systems.[18]

In this paper, we present biomedical researchers' information needs as discovered through user needs analyses and a qualitative usability assessment of DataMed. We share these insights to benefit researchers and system developers working to facilitate biomedical data science, highlighting the importance of understanding biomedical researchers' needs for the success of the systems they build, and suggesting promising areas of future research for the development of successful biomedical data discovery systems. The protocols for our studies are included as appendices to the paper that can be repurposed or used as a starting point for evaluating systems for searching or exploring biomedical data.

MATERIALS AND METHODS

We conducted user analyses in 2 phases to understand researchers' needs and guide the development of DataMed (Figure 1). In Phase 1, we interviewed researchers to understand why and how they look for online biomedical datasets. In Phase 2, we conducted "think-aloud" usability evaluations to assess DataMed's effectiveness at facilitating data discovery.[18]

Potential users of DataMed were defined as researchers involved in biomedical research with experience in biomedical data analysis. Our sampling plan included researchers at various levels of expertise (graduate students, postdoctoral researchers, and faculty members) and various research domains to capture broad patterns of data discovery. Demographic characteristics such as age, gender, and ethnicity were noted during study sessions but not used as selection criteria. Our study protocols were deemed exempt from review by the University of Texas Health Science Center Committee for the Protection of Human Subjects, since we did not collect personally identifiable information and the study did not involve vulnerable populations. Participants were compensated for their time.

Phase 1: user needs analysis

Participants in the user needs study were recruited by an e-mail sent to various universities affiliated with the BD2K group and Texas Medical Center. Thirteen researchers responded to our call and were approved for the user needs analysis (Table 1).

Author VN conducted interviews, in person or remotely depending on the participant's location and preference, lasting from 30 min to an hour. Interview questions were developed by authors VN and TJ as prompts to discuss researchers' current data discovery practices and to identify user needs for a DDI (study protocol available in Appendix Phase 1). During the interview, participants were introduced to the bioCADDIE project and were asked to describe their research area and their experience with 4 aspects of data discovery: searching for data, metadata, data formats, and data visualization. Additional questions were asked to clarify and probe topics brought up by participants.

Author VN took detailed notes on responses to each aspect of data discovery, and then coded them manually using standard word processing software to identify existing practices for data discovery, challenges or areas of difficulty, and design ideas. The coded interviews were summarized across participants to inform the DataMed development.

Table 1. Characteristics of participants in the Phase 1 user needs analysis for DataMed

Research Domain	Position	Count
Clinical Translational Science	Professor	1
Cardiology	Professor	1
Genomics	Postdoctoral Researcher	1
Biomedical Informatics	Professor	4
	Postdoctoral Researcher	1
Molecular Biology	Postdoctoral Researcher	1
Neuroscience	Professor	1
Mobile Health	Postdoctoral Researcher	1
Public Health	PhD Student	1
Anesthesiology	Professor	1
Total		13

Table 2. Characteristics of participants in Phase 2 usability evaluation of DataMed versions 0.5 and 2.0

DataMed Version	Research Domain	Position	Count
0.5	Molecular Biology	Postdoctoral Researcher	2
		Data-related Professional	1
		PhD Student	1
	Chemistry	Professor	1
	Biomedical Informatics	PhD Student	1
	Library Science	Data-related Professional	2
	Total (Version 0.5)		8
2.0	Cancer Biology and Genetics	MD, PhD Student	1
		PhD Student	1
	Cancer Genomics	PhD Student	1
	Public Health	Professor	1
		PhD Student	1
	Genetic Epidemiology	Professor	1
	Systems Biology	Postdoctoral Researcher	2
	Data Curation	Data-related Professional	1
	Medical Library	Medical Librarian	1
	Neuroscience	Postdoctoral Researcher	1
	Total (Version 2.0)		11

Phase 2: usability evaluations

We conducted usability evaluations of DataMed versions 0.5 and 2.0 following their releases. Participants were recruited by e-mail and through flyers from universities affiliated with BD2K projects. Interested researchers filled out a survey sharing their research area and prior experience with DataMed to ensure that those who had taken part in previous studies or had previously used DataMed were excluded. Eight qualified researchers responded to the call for the DataMed 0.5 study and 11 for the DataMed 2.0 study (Table 2).

Authors MS and RD conducted the moderated "think-aloud" usability tests for versions 0.5 and 2.0, respectively.

The semistructured usability test plans (see Appendix Phase II) were developed by each author along with feedback from TJ to simulate representative tasks on the interface and gather feedback on specific design features. The sessions lasted 1 h either in person or remotely, depending on the participant's location and preference. Participants were given a brief introduction to DataMed, then asked to search for datasets related to their research area or interest, stopping when they had found relevant data or could not proceed further. Additional questions were asked to clarify and probe specific topics or issues. Finally, participants were asked to complete a standard usability questionnaire for DataMed 0.5, System Usability Scale for DataMed 2.0, and answer open-ended reflection questions about their experience with DataMed.[19,20]

Both authors took detailed notes on participants' use of the system and feedback during each session; the notes for each session were coded using standard word processing software to identify participants' information needs while exploring DataMed, as well as usability issues that arose during their use of DataMed and suggestions for improvement. These categories were synthesized into recommendations for DataMed's features and user interface. To better understand participants' data discovery processes, several trade-offs were made in the analysis. Given the formative and exploratory nature of these studies, we focused on a detailed qualitative analysis of volunteer researchers. This allowed us to comprehensively analyze participants' interaction with the system; however, this also meant that the quantitative questionnaire results lacked statistical power and did not play much of a role in our analysis. For these reasons, we have omitted the questionnaire results and focus on the qualitative insights from these studies in the following section. Additionally, many participants were from the biological sciences (10 out of 19), which affects the representativeness of our findings for other fields of biomedicine. However, this also reflects the distribution of data indexed in DataMed and suggests that data-intensive biomedical practices using publicly available data may be skewed toward fields such as molecular biology and genetics. As we report and discuss our qualitative findings, we emphasize the perspectives of participants in the translational, clinical, and public health fields as well.

RESULTS

Phase 1: user needs in biomedical data discovery

Participants reported significant effort and difficulty in finding and evaluating relevant data online. One informatics researcher mentioned, "*For all of [our] studies, we would like to integrate relevant studies from other sources. One challenge is knowing what data out there is relevant to what we are doing.*" Common reasons for searching online for data included to validate their own work (such as the effect of an intervention on a cell line), enrich or guide their analyses (such as translational research combining -omics and clinical data), or access data they could not generate on their own (such as hospital data). While many had established strategies such as searching Google or specific data repositories, they found it challenging to know what potentially relevant data was available.

A common frustration was the lack of information in metadata describing datasets. Metadata often contain only partial descriptions of crucial information, such as the samples and techniques used in generating the data. One molecular biology researcher commented, "*The metadata that I would like but usually don't get from my metadata sources is a clean description of tissues that the experimental results came from, the condition of that tissue, the methods of analysis for the phenotypes of interest that were being studied when the experiment was done.*" Additionally, the variety of terminologies used to describe data, the lack of definitions, and poor documentation about the context of the data collection made it difficult to assess its potential usefulness. In these situations, researchers either had to download the dataset to inspect its contents or forgo using it altogether. Further, accessing data through various sources was often fraught with poor documentation of processes required to

Table 3. Summary of user needs analysis for biomedical data discovery

Topic	Difficulties	User Needs
Searching for Data	Time and effort spent finding relevant data for research purposes	Centralized source for available data and tools for finding research-related data
	Poor documentation and protocols for accessing data	Standard documentation and protocols for data access
Metadata	Assessing validity and utility of dataset for secondary use	Standard metadata, vocabularies, and documentation of datasets
	Incomplete, inconsistent, and poor-quality metadata	Tools and guidelines for authors to create metadata
Data Format	Data wrangling and compatibility with analytic methods	Documentation of data provenance
	Availability of data at various degrees of processing: raw to summarized	Availability of data for compatibility with analytic tools
Visualization	Manual work required for creating custom overviews of data	Online visualization of datasets
	Limitation of current methods for visualizing and exploring large datasets	New techniques for representing and exploring large datasets

download the data. This was especially troublesome for those interested in clinical data due to ambiguous institutional review board approval processes. As a neuroscience researcher put it, "*We usually know what we want, but to get access is really like diving into the ocean and trying to reach the other side of the world.*"

The variability of data formats and levels of processing also required significant work to wrangle the data into formats compatible with their processes and tools. An example given by a neuroscience researcher about the limitations of data archives was, "*If you had a collection of image data and genetics… and you wanted to say, find me all the people who have this particular brain difference and also have these two SNPs, you can't do that at all.*" As with metadata, standard information and proper documentation of how the data was generated – its provenance – was necessary for evaluating the potential utility of a dataset. Participants also recognized visualization as a useful way to provide an overview of datasets, variables, and analytic results. However, few online data sources provided such visualization tools, and current visualization techniques were often not flexible or scalable enough to meet their needs, often requiring costly investment in custom visualizations.

Our analysis, summarized in Table 3, validated the concept of a DDI and highlighted key issues with metadata, data standards, and visualization in discovering biomedical data. These findings indicated that researchers would benefit from a centralized source and complete metadata documentation for finding and assessing potentially relevant datasets. Additionally, they need clear protocols to download the data, the ability to download in multiple formats, and a means to visually explore datasets.

These results informed the initial development of DataMed. While researchers currently use generic search engines such as Google, DataMed is intended to aggregate metadata across multiple repositories to provide both broad coverage and effective data-specific retrieval tools, accelerating researchers' exploration of and exposure to potentially relevant data.[21] The DATS model was implemented as a common metadata standard to address the disparity of metadata across repositories. Information retrieval techniques were applied to address the variability in terminology and provide an efficient user interface.[9]

Phase 2: DataMed evaluation

This section describes the combined results of usability evaluations of DataMed v0.5 and 2.0 (Figure 2) following their public release.

Participants encountered difficulty generating queries in the system to describe their information needs. While the search bar on the homepage suggested an intuitive search interaction like PubMed or Google, it was not clear how this interaction would work for complex queries. One researcher commented, "*I would love to search for phenotypes.… For instance, you could search for headache, but I'm not just interested in headache, I'm interested in genes.… How do you search for both?*" Participants' information needs were thus multidimensional, layered depending on whether they had specific research questions they wanted to find data about or were exploring what datasets were available in a domain or for an analytic technique.

Participants also faced difficulties assessing the relevance of datasets returned in DataMed (Figure 3). The most significant problem in assessing the utility of a dataset was inconsistent, incomplete, and poor-quality metadata. Upon searching for a specific cancer, one researcher said of a returned result, "*This doesn't give you any information.… I wouldn't even know what this is about at all.… What does that mean?*" Participants looking for combinations of biomedical concepts and data provenance items found the information about most datasets in DataMed insufficient to determine whether they would be useful. When asked what metadata they would need to evaluate the relevance of a dataset, participants mentioned items included in their information needs, and also added characteristics such as ownership, research organization, and publications based on the data. Common metadata needs across participants are summarized in Table 4.

Overall, while participants found the concept of a DDI valuable, they faced difficulty in understanding the scope of DataMed; it was not immediately clear to them who or what DataMed was intended for and whether it contained data relevant for their research topics. One public health researcher's initial thoughts upon seeing DataMed were: "*This makes me think it's more of a bioinformatics type thing.… The numbers make me scared.… Can I do this, can I not do this?… Do I need a person from a bioinformatics side?*" Users were confused about whether returned results were data or publications ("*I don't know what I'm looking at.… [Is this] a paper, or a grant, or a project?*"). Additionally, information retrieval tools such as query expansion, faceted filtering, and advanced search did not support participants as they explored results due to metadata inconsistency across datasets in indexed repositories. Finally, even when potentially relevant search results were identified, participants had to investigate related publications, the dataset repository's data description, or the data itself to gather additional information or understand the terminology necessary to determine its relevance.

Major suggestions for improving DataMed included embedding domain knowledge and concepts into the organization of the system.

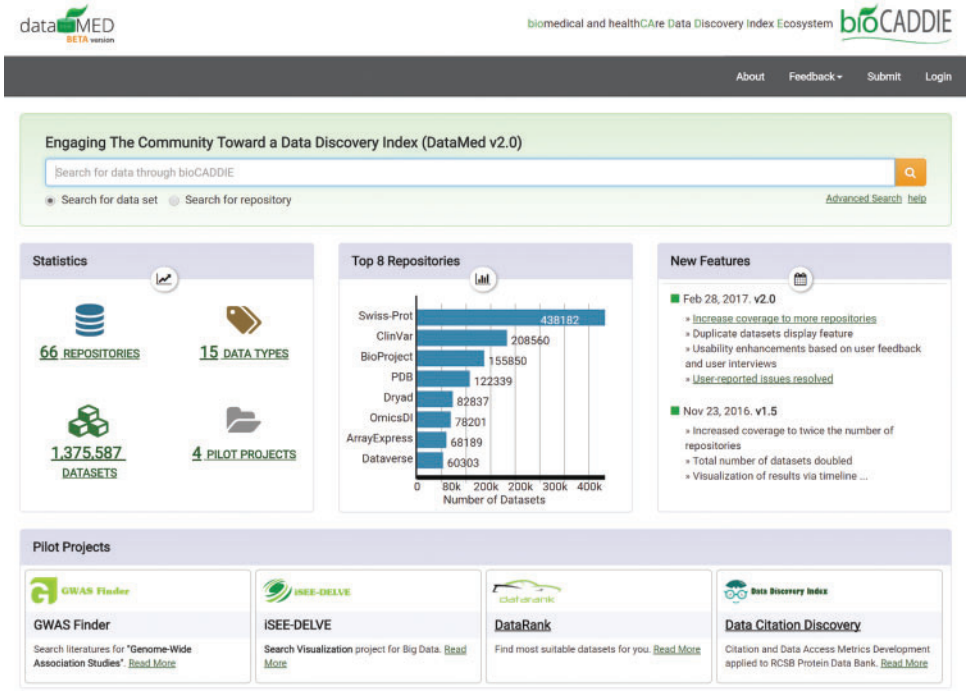

Figure 2. The homepage of DataMed version 2.0 as of May 8, 2017.

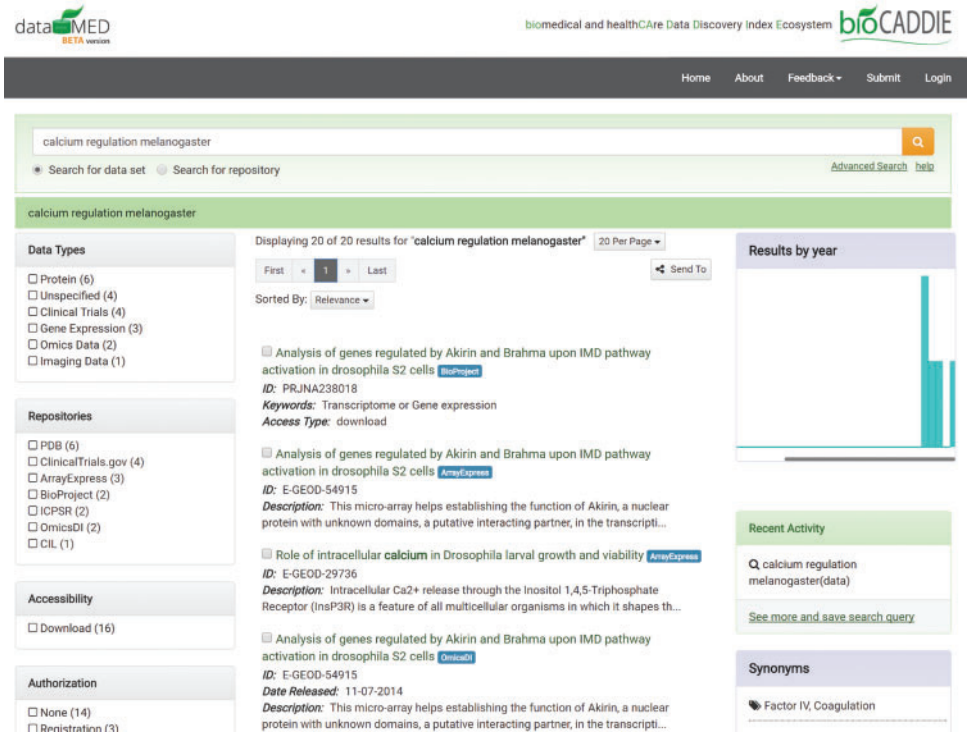

Figure 3. An example of search results for the query "MRI patients Parkinsons" in DataMed version 2.0.

Many suggested providing support for query generation, through interactive ontologies or a "conceptual map," to help them understand how DataMed works and express their information needs more explicitly. They also suggested organizing results by biomedical concepts to provide an overview of returned results, expose DataMed's search process, and improve the utility of faceted filtering for narrowing down relevant results. They suggested more consistent navigation, menus, and customization of metadata fields to improve navigation. Preliminary analyses or summaries of the data itself were mentioned as potentially useful in identifying what findings or associations were discovered in the data and what it could be used for. Finally, researchers appreciated links to other

Table 4. Summary and examples of participants' expressed metadata needs in searching DataMed

Metadata Field	Examples
Biomedical Concepts	De Novo Acute Myeloid Leukemia
Data Type	Gene Expression, Clinical Outcomes
Data Collection Technique	Survey, Magnetic Resonance Imaging
Data Format	Text, Comma-separated Values, Digital Imaging and Communications in Medicine
Data Processing	Raw Data, Abstracted Data, Secondary Data
Sample Description	Number of Samples, Species, Population
Intervention/Study Design	Case-Control, Cohort
Date of Collection	January 2010 to January 2015
Variables	Cell Lines, Hormone Levels, Gene Knockouts
Instructions for Data Usage	Data Processing Tools, Algorithms, Tutorials
Permissions and Ownership	Protected Health Information, Institutional Review Board, Commercial or Academic Research
Research Organization and Principal Investigator	University, Private Institute, International Data
Publications Based on Data	Citations, Papers, Related Items

relevant data or publications that indicated other researchers' use of the data and its utility.

DISCUSSION

Our study provides unique insight into researchers' needs in biomedical data discovery through the development and evaluation of DataMed. While aggregating metadata into a common search engine and applying information retrieval tools is a promising first step, there remain challenges for researchers in finding useful data due to the complexity of data discovery, the lack of metadata standards, and variability in needs across domains and levels of expertise.[2–4] Here we identify these challenges and suggest promising areas of future research for information retrieval systems to support data discovery.

Challenges for information retrieval in data discovery

Supporting exploratory search

Search and discovery in open-ended information systems such as DataMed have no predefined goal; rather, they are exploratory processes motivated by complex information needs directed toward items with characteristics that may or may not exist in the system.[4,12,22] While DataMed provides users with access to a vast number of datasets, the heterogeneity of the information space and ineffectiveness of retrieval tools makes it difficult for researchers to find and evaluate relevant items.[5,11,13,22] These results corroborate previous work and highlight gaps in current techniques for supporting exploratory search.[4,22] Specifically, while user interface elements such as the initial search bar are seemingly intuitive, researchers were uncertain about what they were searching, what constraints they could enter in the search field, and the degree of specificity needed to express their information needs. Additionally, they were unable to adjust their query or search strategy based on the returned results due to the unclear presentation of results, their perceived ineffectiveness of retrieval methods, and the opaqueness of the search process.[22] Indexing systems and user interfaces that organize the complex multidimensional space of biomedical data and support users in navigating it are interrelated technical and design challenges for DDIs.

Evaluating data as an information resource

Our study of DataMed also shows that the relevance of a dataset as an information resource cannot be determined solely from summaries such as keywords, title, or an abstract, as has been suggested may be the case for publications or websites.[4,15] Evaluating the contents and utility of a dataset requires significant context about the data's provenance – its original purpose, characteristics, processing history, and insights derived – to determine whether it could be reused in a different context.[23] Researchers looking for combinations of biomedical concepts and data characteristics in DataMed results encountered difficulty due to incomplete metadata (missing information), poor-quality metadata (incoherent descriptions), and inconsistency in metadata (available fields and terminology) across datasets. Metadata issues also limited the effectiveness of query expansion, faceted filtering, and advanced search features in providing researchers with overviews and navigation tools for identifying and navigating to items of interest. Thus, adherence to standards for metadata representation and quality is another challenge DDIs must address.

Variability across research domains and levels of expertise

Participants' experience and information needs in DataMed also varied with their research domain and level of domain and technical expertise. Novices in technical domains were intimidated by the amounts of data available, while domain experts mentioned that the interface should be more intelligent to understand their information needs. Metadata issues were common across all participants: even advanced researchers could not interpret datasets that had poor metadata. Supporting users from a variety of research and technical backgrounds and with a variety of purposes is an ongoing challenge for any DDI.

Future research areas for biomedical data discovery

Structured metadata abstract

Our study of DataMed shows us that making data available is not enough – researchers must be able to easily evaluate the relevance of a dataset and access the data.[5,7,13] Consistent, complete, and high-quality metadata is vital to any data discovery system, and our documentation of researchers' common metadata needs outlines essential information for effective data discovery. To further this area of research and support the development of effective discovery systems, a common model for describing and evaluating data as an information resource (eg, a "structured data abstract") needs to be used. A structured data abstract containing common metadata fields such as those we have documented in this study would facilitate data discovery by providing consistent descriptions of the characteristics and

context of data to enable the effective use of information retrieval tools and allow researchers to easily evaluate datasets for their utility. Similar efforts in bioscience, such as the Information Sharing Architecture framework, are converging on a common model[24]; such models must be expanded to encompass all domains of biomedicine and refined to enable complex human and computer interpretation.[25] The DATS metadata model[10] is a first step toward this goal, and has evolved considerably since the first prototypes of DataMed were released.

Information retrieval for discovery

Researchers' difficulty in using DataMed reflects a mismatch between current retrieval methods based on specific terms and researchers' needs related to biological or data concepts. These findings corroborate previous work regarding open-ended information systems and have similar implications for the development of new methods and user interfaces for exploratory search.[12,13,22] Traditional search methods such as document-level keyword matching do not adequately provide relevant results for datasets – the information participants are looking for about data are more complex and may be distributed across multiple metadata fields, datasets, or associated publications. Researchers' suggestion to embed biomedical concepts into DataMed is akin to using exploratory search guidelines to leverage the semantics of indexed items to organize search and allow conceptual navigation and discovery in the context of biomedical knowledge.[22] Developing retrieval methods that more closely match the semantics of users' information needs is necessary to facilitate data discovery.

User interface design and visualization

In addition to retrieval techniques, another challenge is designing multifaceted, expressive, and concept-based interfaces that allow users with varying backgrounds to learn from interactions and form a clear mental model of the system. Findings from this study and previous work indicate that support for and control over query expressiveness, transparency about the search process and the format of the results, and guidance on search strategies that provide an overview of search results and foster exploratory behavior can support discovery in open-ended information systems.[12,22] Visualization techniques also have the potential to support navigation and analysis of datasets, but current methods do not support this kind of interaction at scale and must be enhanced to support data discovery and dataset exploration.[7] The design space for dynamic search interfaces that incorporate domain knowledge and biomedical concepts to support exploration of datasets is an exciting area for future research.

Evaluation of open-ended discovery systems

Finally, our study is limited in its short-term evaluation of researchers' discovery practices, the small sample of biomedical research domains, and the evaluation of early-stage prototypes of DataMed. While our qualitative and semistructured evaluations of DataMed helped us identify crucial challenges for data discovery, these methods are limited in scope and scale, capturing rich details, high-level descriptions, and short-term interactions with only a few individuals who may or may not be good representatives of the biomedical science community. Data discovery is a continuous process unfolding over time in research contexts; more study is needed to understand data-intensive research practices across domains of biomedicine and to clarify notions of "discoverability" and "relevance." New quantitative or formal methods must be developed to investigate the evolution of data discovery and to complement qualitative methods for analyzing users' search behaviors and system performance in the user-centered design process for DDIs.[4,7]

CONCLUSION

This paper presents findings from a user needs analysis and usability evaluations conducted during the development of DataMed. The user needs analysis validated the need for a DDI and highlighted issues of metadata, data standards, and visualization in data discovery. Usability evaluations of DataMed further validated the concept of a DDI, provided insight into researchers' information needs, and identified unique challenges in designing a DDI. Making data available is not enough to facilitate data discovery: new information retrieval techniques and user interfaces are necessary for dataset exploration. Consistent, complete, and high-quality metadata is vital to enable this process. We emphasize the importance of understanding researchers' information needs in designing data infrastructures to support biomedical data discovery. While available data and researchers' information needs are complex and heterogeneous, a successful DDI must meet these needs and fit the processes of biomedical researchers.

FUNDING

This work was supported by the National Institutes of Health, grant number U24AI117966.

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

LO-M and HX provided the overall conceptual design for DataMed, including the initial target user groups. HX also identified the scope and user groups for the user studies reported here and participated in the design of the study protocols. TRJ, in collaboration with DR, AG, RD, VN, and MS, designed the detailed protocols and oversaw data collection and evaluation. VN, RD, and MS collected data from the study participants. RD, in collaboration with TRJ, wrote the first draft, incorporating a draft on the first study written by VN. All authors reviewed and offered critical comments and/or revisions of the final version and approved its publication.

ACKNOWLEDGMENTS

We thank participants and members of the bioCADDIE Working Group 9 committee for their feedback.

REFERENCES

1. Margolis R, Derr L, Dunn M, *et al*. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. *J Am Med Inform Assoc*. 2014;21(6):957–58.
2. Brennan PF. Crafting the third century of the National Library of Medicine. *J Am Med Inform Assoc*. 2016;23(5):858.
3. Frankel F, Reid R. Big data: distilling meaning from data. *Nature*. 2008;455(7209):30.
4. Bartlett JC, Toms EG. Developing a protocol for bioinformatics analysis: an integrated information behavior and task analysis approach. *J Am Soc Inf Sci Technol*. 2005;56(5):469–82.

5. Van Horn JD, Toga AW. Human neuroimaging as a "Big Data" science. *Brain Imaging Behav.* 2014;8(2):323–31.
6. Luo J, Wu M, Gopukumar D, *et al*. Big data application in biomedical research and health care: a literature review. *Biomed Inform Insights*. 2016;8:1–10.
7. Kandel S, Paepcke A, Hellerstein JM, *et al*. Enterprise data analysis and visualization: an interview study. *IEEE Trans Vis Comput Graph*. 2012;18(12):2917–26.
8. Ohno-Machado L. NIH's big data to knowledge initiative and the advancement of biomedical informatics. *J Am Med Inform Assoc*. 2014; 21(2):193.
9. Ohno-Machado L, Sansone S-A, Alter G, *et al*. Finding useful data across multiple biomedical data repositories using DataMed. *Nat Gen*. 2017;49(6):816–19.
10. Sansone S-A, Gonzalez-Beltran A, Rocca-Serra P, *et al*. DATS, the data tag suite to enable discoverability of datasets. *Scientific Data*. 2017;4:170059.
11. Jensen LJ, Saric J, Bork P. Literature mining for the biologist: from information retrieval to biological discovery. *Nat Rev Genet*. 2006;7(2): 119–29.
12. Hill JR. A conceptual framework for understanding information seeking in open-ended information systems. *Educ Technol Res Dev*. 1999;47(1): 5–27.
13. Chilana PK, Fishman E, Geraghty EM, *et al*. Characterizing data discovery and end-user computing needs in clinical translational science. *J Organ End User Comput*. 2011;23(4):17–30.
14. Hersh WR. Information retrieval and digital libraries. In: Shortliffe EH, Cimino JJ, ed. *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*. London: Springer; 2014: 613–41.
15. Cohen T, Roberts K, Gururaj AE, *et al*. A publicly available benchmark for biomedical dataset retrieval: the reference standard for the 2016 bioCADDIE dataset retrieval challenge. *Database*. 2017;2017:bax061.
16. Wiklund ME, Kendler J, Hochberg L, *et al*. Technical basis for user interface design of health IT. Grant/Contract Reports (NIST GCR 15-996). Gaithersburg, MD: National Institute of Standards and Technology. 2015.
17. Zhang J, Walji MF. TURF: toward a unified framework of EHR usability. *J Biomed Inform*. 2011;44(6):1056–67.
18. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. *J Biomed Inform*. 2004;37(1): 56–76.
19. Brooke J. SUS: A quick and dirty usability scale. *Usability Eval Industry*. 1996;189(194):4–7.
20. Lewis JR. IBM computer usability satisfaction questionnaires: psychometric evaluation and instructions for use. *Int J of Hum Comput Interact*. 1995;7(1):57–78.
21. Dragusin R, Petcu P, Lioma C, *et al*. Specialized tools are needed when searching the web for rare disease diagnoses. *Rare Dis*. 2013;1(1):e25001.
22. White RW, Kules B, Drucker SM, *et al*. Supporting exploratory search: introduction. *Commun ACM*. 2006;49(4):36–39.
23. Buneman P, Khanna S, Tan W-C, ed. Data provenance: some basic issues. In: Kapoor S, Prasad S, eds. *FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*. FSTTCS 2000. Lecture Notes in Computer Science, vol. 1974. Berlin, Heidelberg: Springer; 2000.
24. Sansone S-A, Rocca-Serra P, Field D, *et al*. Toward interoperable bioscience data. *Nat Genet*. 2012;44(2):121–26.
25. Musen MA, Bean CA, Cheung KH, *et al*. The center for expanded data annotation and retrieval. *J Am Med Inform Assoc*. 2015;22(6): 1148–52.