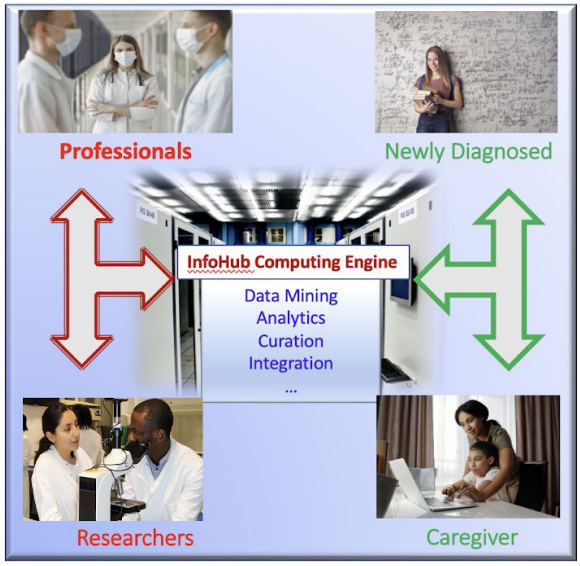
The Rare Disease InfoHub Portal: Actionable Knowledge Appropriately Delivered

David R. Wright, Dave Norris, Alexander Tropsha, Rada Chirkova, Daniel Korn, Jose Quinones, Seth Parrish

The Rare Disease InfoHub (https://rarediseases.oscar.ncsu.edu/) is a web-based platform that provides a one-stop knowledge resource, delivering the latest available information appropriate for patients, clinicians, and researchers alike while providing structural organization and curation to the Rare Disease knowledge continuum. The InfoHub is designed so that as knowledge and computational algorithms advance and new data science tools become available, they can be easily integrated into the platform to take advantage of sources such as:



- Social media streams and related colloquial language analytics;
- Integration of new technologies such as wearable health information and other Internet of Things (IoT) devices;
- Adaptive user interfaces and personalization.

The design philosophy behind the InfoHub is dedicating to ensuring that:

- Curated knowledge resources are delivered in a concise and approachable format to all users;
- Content is research driven, a product of continuously evolving university and global research;
- InfoHub is provided as a non-commercial public service of a nationally-recognized Land Grant University, with no ads and no solicitations for support;
- User privacy is protected, with no collection of personal information or "sign-ups" required to access and of the InfoHub resources.

Introduction

The Rare Disease InfoHub is a key outcome of a multi-disciplinary and multi-institution Research Opportunities for Innovation (ROI) grant from the University of North Carolina System. The InfoHub was created to help improve our understanding of diseases and accelerate the development of novel therapies. The team includes world-class researchers in Data Science as well as domain experts in clinical informatics and drug discovery from the University of North Carolina at Chapel Hill (UNC-CH), North Carolina State University (NCSU), and North Carolina Central University (NCCU). The InfoHub portal is a web-based platform for integrating and disseminating publicly-accessible information and research vital to rare disease patients, caregivers, clinicians, and researchers.

Rare diseases are defined as those that affect fewer than 200,000 (prevalence less than or equal to 67/100,000, US) or fewer than 50,000 (prevalence 1/2000, UK). An ultra-rare disease affects substantially fewer patients, less than or equal to 6,000 (prevalence 2/100,000, US) [6, 7, 10]. According to the National Organization for Rare Disorders (NORD), there are only 250 treatments [4, 11] for the nearly 7,000 rare

disorders that impact an estimated 25-30 million Americans, while the FDA describes over 400 drugs and biologics developed for rare diseases [13].

The Rare Disease InfoHub uniquely builds upon the juxtaposition of two areas of growing strategic importance to North Carolina. In 2015, North Carolina became the first State to pass a law establishing the Advisory Council for Rare Diseases (H.B.823/ S.L. 2015/199). This Advisory Council is housed within UNC-Chapel Hill School of Medicine and is responsible for advising the Governor and the Secretary of NC Department of Health and Human Services (NC DHHS). At the same time, the State has acknowledged recently the importance of Data Science for its economic development [5]. There are ongoing regional developments in basic, clinical, and informatics research in rare diseases, including: (i) the formation of the UNC Catalyst for Rare Disease Drug Discovery in the UNC Eshelman School of Pharmacy [9], (ii) efforts to compile information on rare disease patients within the Carolina Data Warehouse (CDW) and develop tools to enable the use of CDW data for research, including federation with external clinical, demographic and other data; (iii) multiple academic research groups working on rare diseases, and (iv) a growing number of NC companies dedicated to rare diseases.

Background and Related Work

Rare diseases, by their very nature, present many challenges to the full array of stakeholders. Patients face long and frustrating paths to correct diagnoses, difficulties in finding appropriate treatments, and lost opportunities to participate in new therapies and trials that target their specific condition or symptoms. Families and caregivers often struggle to find resources and information to help meet the wide variety of non-clinical needs to maintain the quality of life for their patients (who are often children). Physicians faced with diagnosing and treating such patients may not recognize associations between symptoms and genetic abnormalities or be aware of the latest research results. Researchers investigating genetic mutations and biochemical pathways for potential drug repurposing and other therapeutic interventions may not be aware of related work in the rare disease domain.

All of these represent situations where data exists, but is scattered over multiple, disjoint sources that deliver content in many different formats and interfaces. Many of these sources, such as research publications, are siloed behind very technical interfaces that may be easy for researchers to use but very difficult for novices and non-scientists. Other sources, such as listings of clinical providers and therapists, are publicly available in a variety of formats, but they can be very difficult to relate to specific conditions. Web searches for specific rare diseases, treatments, or specialists can return millions of results, with little indication of which resources are accurate, reliable, and up-to-date.

The tools of data science provide the foundation for the process of collecting, extracting, curating, and analyzing these diverse information resources. The Rare Disease InfoHub addresses a critical knowledge gap by providing a platform to support data science tools for providing knowledge-driven research, treatment, and day-to-day living resources to support the varied needs of the rare disease community in North Carolina and beyond.

The battle against rare diseases is a long and arduous process requiring many coordinated efforts between clinicians, scientists, and patients across the globe. As noted earlier, there are only a few hundred drugs approved, and about 400 total treatments, for a small fraction of the over 7,000 rare diseases. Most of these diseases are thought to be monogenic. If that alone made it simple to generate treatments, this would have happened already, though clearly this is not the case.

Systematically approaching 7,000 rare diseases requires capturing the relevant information in a useful format. The available chemistry and biology data relevant for Rare Disease drug discovery is extremely diffuse, existing in an array of databases (PubChem, ChEMBL, etc.). Currently, there is no exhaustive repository for

rare diseases designed to specifically assist in drug discovery or development efforts. Instead there are some databases that address rare diseases, with links to underlying genetics (e.g. NIH Office of Rare Disease, NORD, OMIM etc.), and others which address foundational support for research and strive to link patients and families to clinical resources (e.g., Globalgenes.org). The US FDA's Rare Disease Repurposing Database (RDRD) [14] consists of Excel tables containing approved orphan drugs. PhRMA recently collated data on >400 treatments in preclinical and clinical phases but this is just a static PDF file and not a queryable database [1].

Clearly, this lack of a definitive rare disease database is inadequate at present and represents a critical need for rare disease researchers. Researchers also require advanced data management facilities for consolidating the underlying genetic and protein causes, and potential treatments of these disparate rare diseases. Bringing them all together in a comprehensive database with information that reaches beyond just the underlying gene will be critical to researchers performing drug discovery on these disparate diseases.

Integrating large amounts of chemical, biological, textual, and clinical data from various databases that use different identifiers, names, or expressions for the same systems represent a real challenge. Beyond the minor but not negligible technical issues mainly related to database (in)compatibilities, a complex workflow has to be developed in order to merge data from different repositories. Figure 1 provides a high-level view of the vision for the Rare Disease InfoHub.

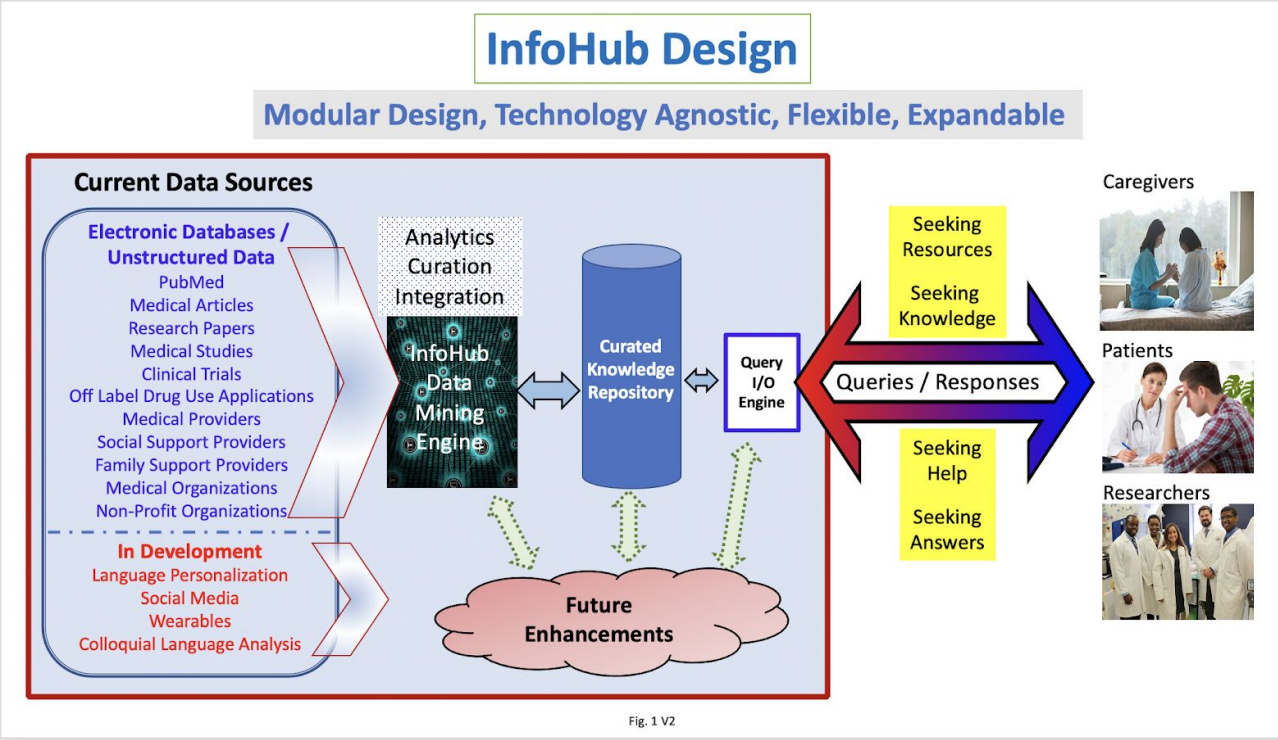


Figure 1. *A high-level view of the vision for the Rare Disease InfoHub, illustrating the application of data science tools to curate diverse information resources, enabling user-sensitive delivery of reliable knowledge resources to support the rare disease community.*

The portal elements include tools for data extraction from primary sources; initial data integration and curation into a structured repository. The expected users represent the breadth of the rare disease community, from patients and their immediate circle of caregivers, through clinical and therapeutic providers, to clinical and laboratory researchers.

Current State

The vision of the Rare Disease InfoHub is that of a unique public portal providing rare disease patients and their caregivers with personalized, time-sensitive, and accurate information about available resources concerning both medical care and quality of life (QoL) support. The InfoHub is a single source of reliable, actionable knowledge resources that:

- supports the decisions affecting the quality of life for Patients and their Families;
- guides the treatments, therapies, and services rendered by Physicians, Therapists, and Caregivers;
- generates insights for Rare Disease Researchers; and
- unifies access to critical resources for the entire Rare Disease Community.

The Rare Disease InfoHub was started as part of a Research Opportunities Initiative (ROI) grant from the UNC System and funded by the NC General Assembly. The InfoHub delivers actionable knowledge resources for patients and their families, caregivers, and researchers through a unique Web-based analytics platform. The proposed center's vision and mission align with NC State's general mission by:

- Addressing very significant challenges in Rare Disease research and care delivery by enabling innovative research in drug and therapy development and repurposing that touch lives in North Carolina and worldwide;
- Engaging with pharma and healthcare industry partners in research and clinical evaluations to identify and evaluate treatments and therapies for new diseases;
- Building partnerships with patient groups and non-profit advocacy organizations to develop and deliver reliable and useful knowledge resources that directly impact the quality of life for rare disease patients and their families;
- Delivering a unique, single point of access knowledge resource for the Rare Disease Community.

Figure 2 shows a high-level view of the current InfoHub software architecture. Behind the scenes, a collection of data acquisition tools mines various databases, websites, and other online data sources to gather information about rare diseases, research publications, healthcare providers, and more. These diverse data sources are curated and integrated into the InfoHub knowledge base, which is stored locally to provide an efficient and responsive user interface. Again behind the scenes, user requests are processed interactively, marshalling data from the internal data store, formatting it appropriately for the user's request and device, then dynamically generating web content which is then sent back to the user.

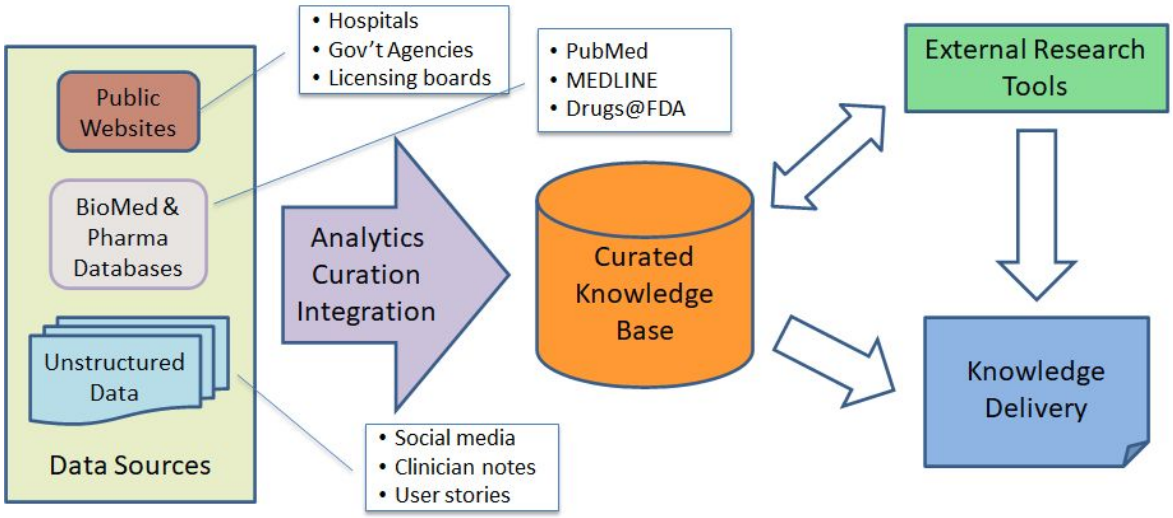


Figure 2. *Abstract view of the current InfoHub implementation*

An important aspect of this architecture is its extensibility, providing the ability for the InfoHub to act as a knowledge source and delivery mechanism for external research tools. The portal incorporates a protected programmatic interface that allows the development and integration of additional research tools access to the curated knowledge base. The InfoHub user interface architecture can also be easily extended to deliver new knowledge resources that are seamlessly integrated into the existing knowledge delivery system.

Extended Research

In parallel with the development of the InfoHub portal, five research projects have been initiated to extend the tools available to the rare disease community as well as to demonstrate the capabilities of using the InfoHub as a research platform. This section briefly outlines these projects.

Radio Transcript Coding

The goal of the Radio Transcript Coding project is to create a verified and validated corpus of colloquial-language texts that have been annotated to label disease, symptom, drug, and related terms and their medically-oriented relationships. The texts under consideration at present are transcriptions of broadcasts of The People's Pharmacy Radio Program [12].

This corpus will provide reliable training and oracle data for machine learning tools that are designed to computationally annotate highly unstructured texts with medically-relevant data and relationships. The resulting annotations can then be used to build knowledge graphs to enable efficient storage and use of the knowledge contained in the texts.

In the short term, the knowledge extracted from these transcripts can be integrated with other knowledge graph sources, e.g., ROBOKOP [3], to help predict previously unknown applications of drugs and other therapies, especially those that may be useful to rare disease patients.

The longer term vision for the outcomes of this project include a tool for analyzing textual narratives submitted by InfoHub users. These narratives could represent real-life experiences of rare disease patients, families, and caregivers. Using analytic tools derived from this project could enable building an interlinked knowledge base to supplement information mined from web-based sources, as well as providing new insights into treatments and therapies derived from actual experience mined from anecdotal evidence on social media and other unstructured text sources.

Knowledge Graph Feature Representation [8]

The goal of this project is to apply novel knowledge graph (KG) representations and machine learning to generate feature representations from existing biomedical knowledge graphs. The resulting feature graphs could then be analyzed to predict previously unknown relationships between drugs, diseases/symptoms, and underlying biochemical processes.

At the core of this work is the need to eliminate the "noise" found in many knowledge graphs. This noise is the result of connections between knowledge elements that are outside the domain of investigation. For example, a knowledge graph built from drug research publications will contain links between drugs and authors, journals, and institutions in addition to the desired connections between drugs, diseases, genes, and biochemical processes. Using the entire knowledge graph becomes very computationally intensive and, in many cases, intractable. Automatically filtering out the undesired noise results in a much more compact data structure that matches the semantics of the domain of investigation, thus improving the efficiency of queries against that data.

This could be a valuable resource for rare disease treatment as it would be an economical means of identifying potential therapies that otherwise might not have been found (essentially hidden in the noise). In the broader vision for the InfoHub portal, the outcomes of this project could be integrated into the KG infrastructure of the next-generation portal.

Knowledge Graph Extraction & Refinement

The knowledge graph (KG) extraction & refinement project aims to develop tools for refining existing knowledge graphs.

The KG extraction part of the project involves developing tools for automatically annotating colloquial text sources and building a KG from the resulting annotations. This part of the project is closely related to the Transcript Coding project, which will provide the annotated texts for training and validation of the resulting extraction tools.

The proposed refinement process includes two parts: KG completion and KG error detection.

KG completion will allow missing knowledge to be added to an existing KG through a combination of

- predicting missing entries;
- predicting missing types/classifications; and
- predicting missing relationships within the KG.

Reliably predicting where knowledge is missing will enable subject matter experts (SMEs) or (possibly) automated tools to efficiently insert the missing elements.

The second part of KG refinement involves detecting errors within a KG. Within a single KG, these errors could be incorrect type associations, relations, or literal values. KG error detection will also identify errors in links between two different KGs.

The outcomes of this project have been documented in a manuscript currently under revision for publication [2]. This work will also serve as a foundation for a new and more efficient KG infrastructure for the InfoHub portal.

Data Curation

The Data Curation project is investigating the theoretical underpinnings of cleaning and repairing data collections with the goal of ensuring data quality and veracity. The goal is to develop a sound basis for developing tools to unify and curate data from multiple sources and implement those tools to work with the disparate information sources present in the rare disease and drug repurposing data flows.

The outcomes of this project will also be fundamental to a new InfoHub KG infrastructure.

Avatar User Interface (AUI)

A continuing criticism of the current Rare Disease InfoHub portal is that it does not invoke a pleasant user experience. While improvements have been made to make the site easier to navigate, it still remains very "functional" and less user-friendly.

The goal of the Avatar User Interface project is to develop a User Interface (UI) that would:

- be simpler to use for all InfoHub Site visitors;
- be attractive and welcoming;
- provide a means for anonymously organizing InfoHub resources based on the user's chosen level of experience and where they are in their 'knowledge journey';
- be extensible to a variety of new resources and features.

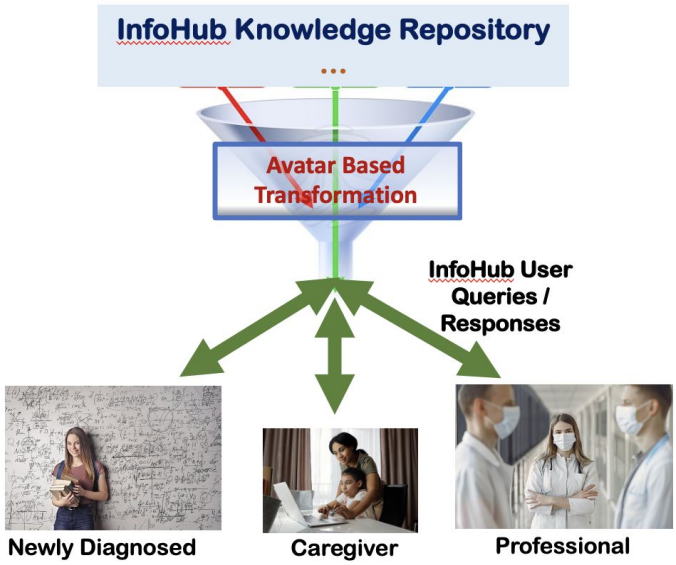
The Avatar User Interface (AUI), is unique in how it advances the delivery of what can be very complex information, to the Rare Disease community, in a 'user centric' personalized way. The design is based on techniques used in Game Theory where a player is given more and more detailed content as they progress through the stages of Game Play. Using the AUI, the InfoHub Repository content is appropriately adjusted and modified to the User's declared level in their Knowledge Journey.

The AUI allows for an InfoHub user to declare themselves at 1 of 3 'knowledge levels':

1. Newly Diagnosed
2. Caregiver
3. Professional

This 'User Knowledge Level Identifier' is then used by the AUI to transform InfoHub Repository content, which is being requested by the User, to their chosen level of knowledge, allowing the system to deliver information using language and terminology based on where the User is in their declared "Knowledge Level".

The system can remember the user settings across site visits (persistence) and the user is able to change their chosen 'Knowledge Level' at any time, as their own personal knowledge and experience grows, thus obtaining more and more detailed information in their query responses from the InfoHub Portal.



User Selected Knowledge Levels

Future Work

Technical Development

Figure 3 illustrates our proposed technical vision for guiding future InfoHub research and development. This vision is intended to provide a much friendlier user experience while facilitating the integration of new and more powerful tools into a user's workflow. The core of this vision is the development of a Knowledge Graph Infrastructure (KGI) within the InfoHub that enables accurate and automatic integration of multiple sources and types of information of interest to Rare Disease patients, families, providers, and researchers. Wrapping this infrastructure will be a completely new kind of user interface designed to effectively deliver this knowledge to users in different ways depending upon their personal needs, experience, and existing knowledge.

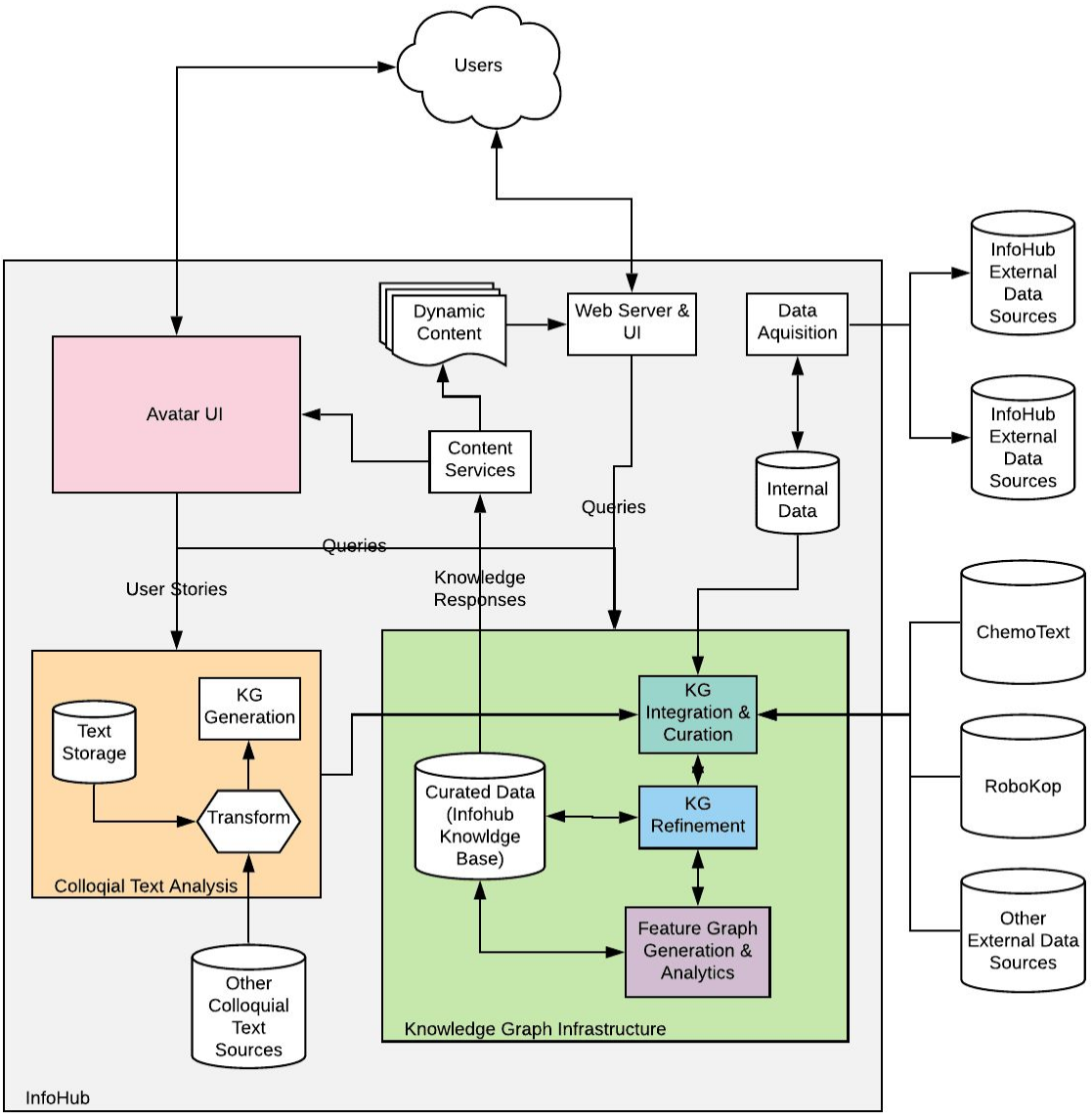


Figure 3. *Future InfoHub Architecture*

In addition to the Avatar User Interface (AUI) goals listed above, the vision for the future of the InfoHub includes providing users the ability to share their experiences (stories) with the community and the organized presentation of these shared community resources to other users while protecting user identities. Following a

validation process by an independent site moderator, user stories will be stored verbatim for presentation to other users as well as provide further analytic processing to extract knowledge elements.

The Colloquial Text Analysis (CTA) module will be a derivative of the Radio Transcript Encoding and Knowledge Graph Extraction research projects described above. The purpose of this module will be to take colloquial text sources and generate knowledge graphs that can be integrated with other knowledge sources into the InfoHub Knowledge Base. While not in highly structured forms (like clinical health records), these user stories can contain valuable experiential knowledge about potential treatments and therapies, community resources, healthcare providers and specialists, and more. Capturing this knowledge and integrating it into the InfoHub will make it readily available to more users.

A critical element of this vision for the future of the Rare Disease InfoHub is the KGI. There are two main steps in the process of building the InfoHub Knowledge Base (IKB). First, incoming information will be integrated into a single store of knowledge that has been curated to help ensure its completeness and correctness. This information will come from a variety of sources, including current InfoHub information, output from the CTA, and information from ChemoText and ROBOKOP. The second key step in the KGI is to refine the InfoHub knowledge base through advanced analytical techniques. This will involve using information from the diverse input sources to identify missing or incorrect data and automatically make valid corrections.

Another feature of the KGI and the IKB is that it will support analytic discovery of potential drug and therapy repurposing. By integrating information and knowledge from multiple disparate sources, new knowledge can be inferred from existing relationships that would have gone unnoticed without the large-scale integration in the IKB. While the primary target of these queries will be in support of new Rare Disease treatments, the applications of the underlying technology can apply to all domains of medicine.

As noted earlier, this architectural vision for the future of the InfoHub will enable better user experiences, advanced analytic capabilities, and provide a consistent structure for anning new information sources, analytic tools, and user interface features.

Community Development

This technical development will greatly enhance the Rare Disease InfoHub's value as a resource to the Rare Disease community. Figure 5 captures a vision of how the InfoHub fits into a larger vision of community-building for patients, caregivers, clinicians, and researchers in the Rare Disease space. The InfoHub is seen as the "hub" in bridging advanced analytics and computational technology with pharma and biomedical informatics research and rare disease patients and community, facilitating the exchange of knowledge and experience between these different but interdependent groups.

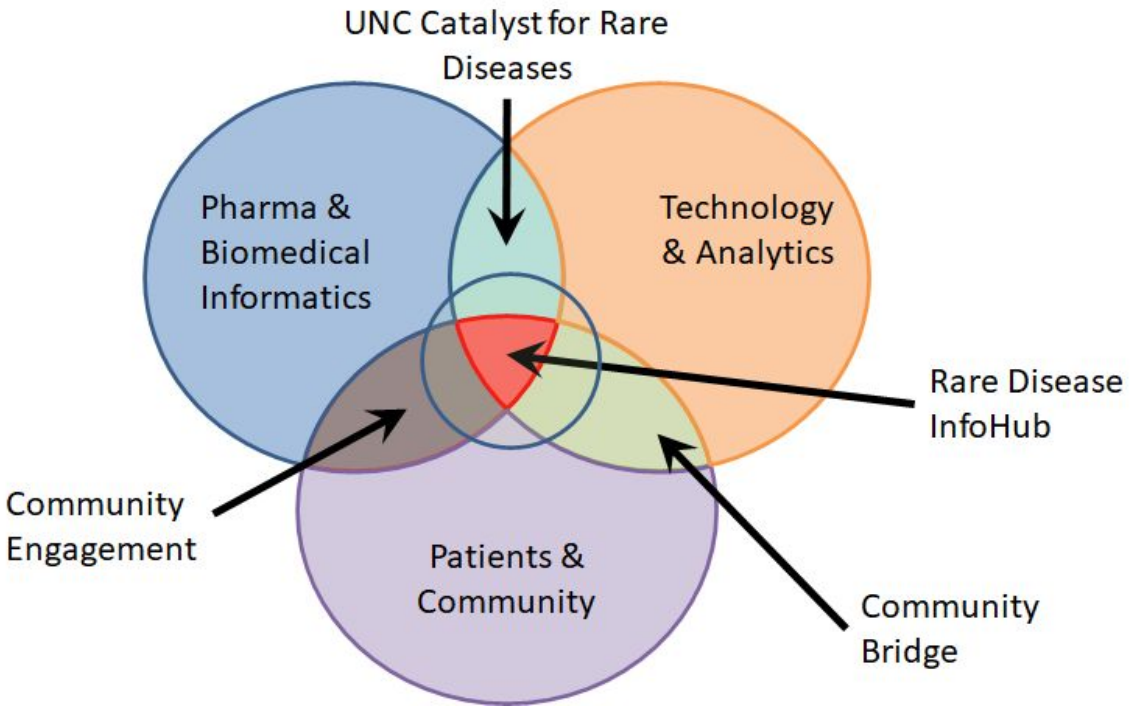


Figure 5. *InfoHub at the center of building Rare Disease Communities*

References

[1] Rare Diseases: A Report on Orphan Drugs in the Pipeline. Anonymous https://www.slideshare.net/PhRMA/rare-diseases-a-report-on-orphan-drugs-in-the-pipeline.

[2] Text Mining to identify and extract novel disease treatments from unstructured datasets. ABRAR, S., YEDIDA, R., TROPSHA, A., CHIRKOVA, R., MELO-FILHO, C., KORN, D. AND MURATOV, E. .

[3] BIZON, C., COX, S., BALHOFF, J., KEBEDE, Y., WANG, P., MORTON, K., FECHO, K. AND TROPSHA, A. 2019. ROBOKOP KG and KGB: Integrated Knowledge Graphs from Federated Sources. *Journal of Chemical Information and Modeling* 59, 4968-4973. https://doi.org/10.1021/acs.jcim.9b00683.

[4] COTÉ, T.R., XU, K. AND PARISER, A.R. 2010. Accelerating orphan drug development. *Nature Reviews Drug Discovery* 9, 901-902. .

[5] North Carolina in the Next Tech Tsunami: Navigating the Data Economy. DORON, S., MCKEEN, S., LEONCHUK, L., KNOWLES, D. AND HARDIN, J. .

[6] What is a rare disease? EURODIS. .

[7] RARE Facts. GLOBAL GENES. https://globalgenes.org/rare-facts/.

[8] Noise Removal with Regular Expression in Knowledge Graphs. HOU, P., CHIRKOVA, R. AND WRIGHT, D. .

[9] UNC Catalyst Initiative Aims to Create and Share Tools to Fight Rare Diseases. MENDENHALL, G. https://pharmacy.unc.edu/2017/01/unc-catalyst-initiative-aims-create-share-tools-fight-rare-diseases/.

[10] List of Rare Disease Information. NORD. https://rarediseases.org/for-patients-and-families/information-resources/rare-disease-information/.

[11] Rare Disease Database. NORD. https://rarediseases.org/.

[12] The People's Pharmacy Radio Program. NORTH CAROLINA PUBLIC RADIO. https://www.npr.org/podcasts/381444414/the-people-s-pharmacy-radio-program.

[13] Developing Products for Rare Diseases & Conditions. OFFICE OF ORPHAN PRODUCTS DEVELOPMENT. https://www.fda.gov/industry/developing-products-rare-diseases-conditions.

[14] XU, K. AND COTE, T.R. 2011. Database identifies FDA-approved drugs with potential to be repurposed for treatment of orphan diseases. *Briefings in bioinformatics* 12, 341-345. .