Exposome, 2023, **3(1)**, osad010
https://doi.org/10.1093/exposome/osad010
Commentary

2022 NIEHS Catalytic Workshop Series on the Exposome

A roadmap to advance exposomics through federation of data

Charles P. Schmitt [1,*], Jeanette A. Stingone[2], Arcot Rajasekar [3,4], Yuxia Cui[5], Xiuxia Du[6], Chris Duncan[7], Michelle Heacock[8], Hui Hu[9], Juan R. Gonzalez[10], Paul D. Juarez[11], Alex I. Smirnov[12]

[1]Office of Data Science, National Institute of Environmental Health Sciences, Durham, NC, USA
[2]Department of Epidemiology, Mailman School of Public Health, Columbia University, New York, NY, USA
[3]Renaissance Computing Institute, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[4]School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[5]Exposure, Response, and Technology Branch, National Institute of Environmental Health Sciences, Durham, NC, USA
[6]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC, USA
[7]Genes, Environment, and Health Branch, National Institute of Environmental Health Sciences, Durham, NC, USA
[8]Hazardous Substances Research Branch, National Institute of Environmental Health Sciences, Durham, NC, USA
[9]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA
[10]Center for Research in Environmental Epidemiology, Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain
[11]Department of Family & Community Medicine, Meharry Medical College, Nashville, TN, USA
[12]Department of Chemistry, North Carolina State University, Raleigh, NC, USA

*To whom correspondence should be addressed: Email: Charles.schmitt@nih.gov

Abstract

The scale of the human exposome, which covers all environmental exposures encountered from conception to death, presents major challenges in managing, sharing, and integrating a myriad of relevant data types and available data sets for the benefit of exposomics research and public health. By addressing these challenges, the exposomics research community will be able to greatly expand on its ability to aggregate study data for new discoveries, construct and update novel exposomics data sets for building artificial intelligence and machine learning-based models, rapidly survey emerging issues, and advance the application of data-driven science. The diversity of the field, which spans multiple subfields of science disciplines and different environmental contexts, necessitates adopting data federation approaches to bridge between numerous geographically and administratively separated data resources that have varying usage, privacy, access, analysis, and discoverability capabilities and constraints. This paper presents use cases, challenges, opportunities, and recommendations for the exposomics community to establish and mature a federated exposomics data ecosystem.

Keywords: data sharing; data ecosystem; data standards; federation; exposome; federated learning

Introduction and vision

The human exposome is the totality of exposures a person receives over their life course and the impact those exposures have on the individual's health. In the Summer and Fall of 2022, the National Institute of Environmental Health Sciences (NIEHS) conducted a series of workshops, called the Catalytic Workshop Series,[1] to understand the current state of exposomics research and to guide future research and investments in enabling capabilities. NIEHS issued ambitious goals for this workshop as "we will explore what it means to conduct exposomics experiments, develop new tools, techniques, and technologies, share data for maximizing in silico experimentation, and cultivate the research continuum from fundamental to population health. All this work will contribute to developing a framework for demonstrating the value of the exposome in environmental health."

During the workshop series, five major areas were identified as necessary to operationalize exposomics. These five areas can be broadly classified as dealing with 'what to measure', 'how to measure', 'share and harmonize', 'integrate, analyze, and interpret' and 'translate and impact.' Several concepts were considered as especially important by participants for near-term focus, including the goal of developing a data ecosystem in which harmonized data can be found, accessed, and shared through sustained and interoperable data repositories. This sentiment is aligned with recent articles[2-4] which have identified a need for integrated solutions for combining internal and external exposomics research, as well as highlighting an urgent need for an information exchange clearinghouse to facilitate sharing of tools and data.

This paper is an outcome of the workshop series discussions with the aim of developing a framework that enhances data exchange for exposomics research. The envisioned data ecosystem framework would facilitate access to and integration of a vast amount of diverse longitudinal data including general external influences (pollution; weather and social context); external

Received: April 17, 2023. **Revised:** September 7, 2023. **Accepted:** October 4, 2023
Published by Oxford University Press 2023.

individual-specific factors (diet, infections, self-selected chemical intake); internal individual-specific constituents (metabolic byproducts, microbiome derivatives, inflammatory mediators, stress hormones, etc.) that contribute to the onset and progression of disease; and measures of personal (behaviors, screening, treatment, outcomes); and population health (disease rates, years of person life lost, longevity) and health care (access, costs, quality). The data ecosystem would enable priority concepts brought forward in the workshops, including focuses on understanding the exposome at a community level, advancing exposome-wide association studies (ExWAS) and functional exposomics, and advancing the research and application of nutritional pharmacology and precision nutrition to modify the effects of exposures. More generally, the data ecosystem would facilitate a broad range of exposome research by ensuring that the data sets needed by researchers meet the FAIR principles, which states that the data should be Findable, Accessible, Interoperable, and Reusable.[5]

We consider that such a data ecosystem should be developed with an international scope enabling sharing of data, tools and workflow pipelines across multi-disciplinary sciences, diversity of analytes, and linking studies conducted across continents and merging data from existing and future longitudinal studies. The data ecosystem would entail a federation of data resources that already exist worldwide and across subfields but requires an adoption of common protocols for data access and sharing, adoption of semantic standards to guide the collection, storage, analysis, and leveraging of data and data resources, such as programming languages, packages, algorithms, and cloud-computing services needed to create and maintain it. The effort can be hastened through development of a consensus, international data governance strategy that adapts guidelines for proposing, adapting, implementing, and evaluating processes (feasibility) and outcomes (high quality, rigorous, and reproducible data). Guidelines are needed for an array of tasks, including adoption of common terminologies and ontologies; identification of data sources; steps for data collection, linkages, transfer, storage, and security; risk management; and addressing confidentiality of protected health information including privacy risks related to geo-spatial data. Much like the Human Genome Project, it will require the establishment of a consortium of transdisciplinary researchers, data stewards and data scientists, and technologists from across the world as well as the support, leadership, and cooperation of funding agencies.

We believe that such an ecosystem will not be bound to a single site or even a single continent. The system would be a federation of federations that can interlink data across multiple sites. Such a loosely coupled structure would help in maintaining autonomy but still be guided by FAIR principles to share across repositories. The need for interoperability across such loose coupling will require a strong set of standardizations which we believe is possible given the early stage of the exposomics field. Institutions such as the HHEAR Data Center[6] and the Exposome Explorer[7] are promoting such common standards for data depositions. Where standardization is not feasible, developing crosswalks would be possible and helpful as we tackle integrating data across multiple disciplines.

Such an effort is tractable and timely. The existing research data ecosystem is used daily by researchers in environmental health to find, access, and work with a diversity of data. Directed efforts are already improving the ecosystem by providing needed funding to sustain data repositories and knowledge bases, to develop standards and related tools, to create libraries of common data elements, and to create common data sharing protocols and technologies. Efforts, such as the European Human Exposome Network (EHEN), the Global Alliance for Genomics and Health (GA4GH), and the NIH Cloud Platform Interoperability (NCPI) project are already developing and providing federation capabilities. The exposome community has the opportunity to work with and build upon these efforts to ensure the emergence of a data ecosystem that serves its use cases and needs.

Workshop participants recommended the creation of an Exposome Community of Practice (CoP) to promote and foster research and impact along multiple avenues, such as supporting workforce development and advancing new data science and statistical methodologies. This paper presents several recommendations, including the utilization of the CoP alongside other community, coordination, and technical activities that aim to evolve the current data ecosystem to the envisioned ecosystem. To inform these recommendations, the paper first presents a high-level overview of the use cases the ecosystem may serve, data collections to consider as foundational, an overview of data federation and its challenges, current opportunities to build upon, and then proceeds with a listing of recommendations.

Driving use cases

The ecosystem should enable use cases that fall within broad categories, including

Pooled analysis

Pooling study data can increase statistical power, as shown in recent multi-cohort analysis to estimate cancer risks due to occupational hexavalent chromium and nickel exposure,[8] air pollution exposure,[9] heat exposure,[10] occupational polycyclic aromatic hydrocarbons exposure,[11] and leisure-time activity.[12] Adoption of standards within the ecosystem will lower barriers for aggregating data and analyzing interactions between multiple environmental factors and covariates and enable future exposomics researchers.[13]

Replicability

Under the RECOVER program, Zhang et al.[14] conducted ExWAS of post-acute sequelae of SARS-CoV-2 infection in INSIGHT (a large New York City clinical research network (CRN)) and replicated findings in OneFlorida+ (a large Florida CRN). They identified air toxicants, particulate matter compositions, neighborhood deprivation, and built environment factors associated with an increased risk of PASC conditions related to nervous, blood, circulatory, endocrine, and other organ systems. Adoption of the FAIR data practices throughout the data ecosystem will foster greater reproducibility and replicability of results and, thus, support translation of the exposomics research into health practices.

Discoverability

Current tools and approaches can excel for such common tasks as finding relevant papers (e.g., Pubmed, web search engines) or discovering new linkages between the existing knowledge (e.g., large language models). These tools, however, are often unsatisfactory for finding resources based on less common but insightful scientific parameters, such as identifying study data by the temporal range and resolution of geospatial exposure estimates used in the study or finding data sets that have similar EWAS signatures to a signature of interest. Adoption of common search methods and standards across the exposomics community could address these gaps and facilitate adoption of new search

approaches (e.g., based on embedding models) for benefits of the field.

Intelligence

Hanson et al.[15] recently published statistics from 10 cohorts showing that the long COVID symptoms are present in 3.69% of infections with fatigue, respiratory, and cognitive symptoms occurring in 51.0%, 60.4%, and 35.4% of cases. Well-defined analysis, such as identifying emerging health concerns, can be executed as periodic and automated cross-repository queries that provide updated reports that inform health and research policies and priorities.

Cohort discovery and recruitment

Systems such as i2b2/Shrine[16] have been providing cohort discovery and recruitment capabilities across data centers for years, such as the NCATS Accrual to Clinical Trials Network of 50+ institutions.[17] The use of federated cohort discovery tools could be applied towards identifying and recruiting exposomics cohorts and identifying replication studies.

Creation of AI/ML data sets

The Barcelona Institute for Global Health recently released an exposomics data set containing multi-omics and multiple exposures and disease phenotype data and conducted and published the results of a data challenge that used state-of-the-art statistical methods for studying exposome-health associations.[18] Such extensive data sets are still a rarity today, but will be increasingly important for advancing exposomics research and for the development of new analytical methods that can deal with high-dimensional and correlated data from internal and external measurements. Realization of recent calls for greater adoption of data-driven approaches in exposomics[19-22] will benefit from a federated ecosystem that allows AI/ML ready data sets to be routinely generated and updated.

Signatures/distributions

Precomputed signatures of omic-related data, such as computed by the BD2K-LINCS[23] aid in the analysis and interpretation of genomic data. Similar signatures can be generated for exposomics. Federation technologies are capable of constructing signatures and background distributions while limiting disclosure of private information,[24] a critical capability given that even moderate precision external exposure data poses re-identification risks.

Alignment

The ability to spatially and temporally align data will be a critical feature of the data ecosystem that allows for identification of sources of exposure and the biological pathways that lead to the onset and progression of disease along the life course.

Expand the visibility and usability of existing data sources

Many data repositories contain data that have been carefully collected in the course of painstaking efforts, but are not widely known and lack modern data visualization and analytics support to make the data easily usable. The federated ecosystem will address these issues. For example, the USDA and CDC websites contain large amounts of food measurement, questionnaire, demographics, and laboratory data that have tremendous value and can benefit the exposomics community much more through a data federation ecosystem.

Foundational data collections

Exposomics incorporates data from many scientific domains, each with its own set of characteristics and minimal information standards and needs to align with the FAIR data principles. Table 1 provides a brief description of diverse data domains that form the basis of exposomics research. The data sources, opportunities, and challenges listed in Table 1 can inform and prioritize efforts to develop foundational data sets, tools, and policies for the ecosystem.

In addition to considering the variety of exposomics data types, a FAIR-aligned data ecosystem requires differentiation between the levels of data that are shared. The NIH Genomics Data Sharing Policy[25] differentiates five-levels of data based on level of processing and aggregation. Level 0 represents raw instrument data, level 1 represents data after initial transformation from the raw format, level 2 is data that has been cleaned and undergone quality assessment, level 3 data has been processed to identify key features, and level 4 data has been integrated with other related information. A similar rubric needs to be applied to exposomics data. For example, ambient air pollution data obtained directly from an air monitor are Level-0 or 1 while -omics data are often Level-3 as these data are typically normalized and annotated. While data at all levels require annotation of the measurement platforms and study characteristics, high levels data also requires information on data processing in order to meet the FAIR standards. Prioritizing the data domains and data levels for sharing will aid the community in allocating the resources available to maintain data repositories and their capabilities.

Federation and the data ecosystem

While federated systems are not new (a NIST reference architecture already exists[26] and multiple guides can be found in popular technical press), prioritizing and tailoring components and capabilities for exposomics research has not been done at the community scale yet. Table 2 provides a listing of important components and capabilities by access type and use case types that are characteristic of a mature federated data ecosystem.

Discovery catalogs

Catalogs that provide up-to-date inventories of relevant resources are needed. Such catalogs can employ search protocols and tools tailored for exposomics (e.g., to limit search to specific exposure media), support search for uncommon data (e.g., specific polyexposure risks), and employ the use of advanced language models (such as the OpenAI GTP systems[27]) to support customized concept-based and natural language search.

Common language

A common language, building off related standards (e.g., Fast Healthcare Interoperability Resources (FHIR)), that covers data structures, formats, representation, as well as data and metadata terminology is a necessary step in creating a federated data ecosystem. Juarez et al.[28] previously identified five broad domains of exposures and subdomains that a common language should incorporate: 1) natural (air, water, soil, and land); 2) built (places you live, work, play and pray); 3) social (social, demographic, economic, and political); 4) policy (federal, state, and local), and 5) health and health care (personal and population health; facilities, finances, and providers). Common metadata elements include spatial and temporal units, source, coding for missing data, web address, and data limitations. Other elements, such as

Table 1. Potential data domains to provide a foundation for a federated data ecosystem

Domain	Example data sources	Example opportunities	Example challenges
Geospatial	Census; neighborhood-level characteristics; land cover/ land use; personal sensors	Expand inclusion of social and structural drivers of exposure and health; development and application of more advanced exposure assessment tools	Privacy and confidentiality of PHI; aligning across varying levels of spatial and temporal aggregation
Omics	Metabolomics; proteomics; genomics	Characterize pathways linking exposure and health; develop new biomarkers of both exposure and response	Documentation and alignment of processing pipelines; high-dimensionality; analysis and transfer of large data, potential sensitivity to processing parameters
Epidemiologic	Questionnaire data; physical and mental assessments; public health monitoring	Expand reuse of existing cohorts; provide data for analytic methods development and pooled projects	Privacy and confidentiality of PHI, use of non-standard measurement tools; lack of a common repository
Environmental monitoring	Ambient air pollution levels; water quality; population-based biomonitoring	Consistent exposure assessment across diverse areas	Calibration across monitoring networks; equity of monitoring placement; data storage
Administrative, clinical, medical records	Billing records; Electronic health records; public health registries (e.g. CDC Wonder, SEER); insurance claims	Efficient construction of large study populations; increased ability to investigate rare diseases and outcomes	Variability in data quality; privacy and confidentiality of phi; variability in consents and use agreements; alignment and harmonization across systems
Foodome and drugome	Nutritional composition of various foods; chemical composition of various drugs; health impact of various nutrients and drug chemicals	Enable studies of how diet and drug can affect human health	Linking nutrients and drugs to metabolic pathways and various diseases
Toxicology and chemical	Toxicology assays on in-vitro and in-vivo systems; predictions of toxicology; chemical toxicology, biological, and health-related annotations	Greater integration of toxicology/chemical knowledge with clinical knowledge sources; incorporation of predictive toxicology with public health planning and interventions	Data standards for sharing; linkage to clinical/biological data; access to industry data

This is not an exhaustive list of resources and databases for exposomics.

Table 2. Key components of a federated data ecosystem for exposomics

Key components	Data types	Important use cases	Description
Discovery catalogs	UR, R	Search	Provide services for finding and discovery of federated resources (data, tools, and cohorts)
Common Language	UR, R	All use cases	Provides standards for communications between resource providers
Access & Rights	R	Pooled Analysis, Intelligence, Cohort Discovery, Signatures, AI/ML data sets	Controlled access to federated resources, including usage rights and restrictions
Federated Analysis	R	Pooled Analysis, Intelligence, Signatures	For analysis of sensitive data in-place
Data Workbenches	UR, R, L, S, C	Pooled Analysis, Intelligence, Signatures, AI/ML data sets	Exposomics specific tools and systems for large and/or restricted data

UR, unrestricted data; R, restricted access data; L, large scale data; S, streaming data; C, multidimensional, complex data.

providing standardized geocoded addresses, will make it easier for end users without creating an undue burden on data generators and repositories.

Access & rights

Controlled access to data is a foundational part of federated systems. This includes federated authorization and authentication mechanisms as well as mechanisms for assigning permissions and restricting access and rights based on data usage agreement and consent agreements. Machine actionable data rights and consents are needed to support efficiency and automation. Tailoring evolving solutions in this space for adoption by resource providers and addressing challenges (e.g., providing common consent language for collecting and linking geospatial data elements) should be a priority.

Federated Analysis

The ability to conduct cross-study analysis without moving sensitive data opens the door to larger pooled data sets that can be re-analyzed frequently as new studies are developed and new

data collected. This not only supports new discoveries but can promote cohort discovery, intelligence, and signature development. The capability requires an agreement on protocols, language, compatible tools, and sustained infrastructure. This is especially important for the exposomics community given the re-identification risk and misuse potential from external exposure data coupled with the needs of precision exposomics to collect high precision and frequent external exposure measures.[29,30]

Data workbenches

Exposomics data can include genomics, metabolomics, epigenomics, clinical, social, and environmental data. This presents significant challenges to research groups that need access to data, data type specific tools, expertise with tools, and the compute infrastructure to work with the data. Often, data will have to be provisioned in secure enclaves that disallow download of data. The exposomics field will benefit from workbenches tailored towards exposomics analysis concepts[2] that data providers and aggregators can readily deploy at low costs (e.g., in cloud).

Data repositories and secure enclaves

While several domain-specific, general-purpose, and study/project specific data repositories do exist, exposomics data is often split between different repositories (e.g., genomic data in GEO, metabolomics data in MetaboLights or Metabolomics Workbench) and study data often resides in (often term-limited) study-specific repositories. Additional repositories are needed to provide for long-term management of exposomics study data and to support the aggregation and linkage of data into AI/ML ready data sets. Data repositories that house sensitive data often provide a secure enclave that allows users to log into a virtual computer to access data without the ability to remove the sensitive data from the enclave. Dashboards that provide data analysis and exploration capabilities while preventing the downloading of the data may also be employed. Providing secure enclaves alongside data repositories will serve to increase access to sensitive data sets and will be important in expanding research that includes sensitive elements such as occupation and geo-coordinates.

Addressing the challenge of federation

Federation is known to be challenging[31-33] as it requires coordination and agreement between different resource providers along multiple dimensions. Providers must adopt common protocols for communications between software systems; common authentication and authorization mechanisms for controlling access to and use of shared resources (including data use agreements and consent agreements); common standards for data and metadata, as well as common protocols for search, retrieval, and cross-platform analysis.

Ensuring operations and maintenance of necessary federation infrastructure is not easy. Computer and network failures can hamper federated services and software typically needs to be continuously updated for new requirements and to deal with changing security issues. Migrating legacy systems and sustaining the budget needed for continued operations is a challenge especially when time-limited funding mechanisms (e.g., grants) are used and where federation goals go beyond the core mission of the funded resources. For example, work to federate exposome and omic data as part of the HELIX project[34] was hampered by mismatches between available server infrastructure and project needs. Cloud resources can help in providing appropriate IT resources as well as in addressing security requirements. Recognizing the design and budget requirement on federation of IT infrastructure will be important for success.

The diversity and breadth of the field of exposomics magnifies these other challenges. Exposomics crosses multiple fields and data types, each having their own standards and tools, which can lead to numerous incompatibility issues and a need for interoperability across standards. Diverse fields such as medicine, environmental sciences, and geosciences do not frequently collaborate, and different funding sources and priorities can hamper such collaborations. The exposomics community is international which brings challenges due to different policies and laws, especially around privacy, as well as practical challenges in coordination.

Federation, however, is not a one-size-fit-all concept and there is flexibility in addressing these challenges. The US Government Data Federation website (https://federation.data.gov/) defines a data federation project as one "in which a common type of data is collected or exchanged across complex, disparate organizational boundaries." Data can be exchanged across organizational boundaries without resolving issues of authentication, harmonization can be conducted after exchange of data or limited to data where aggregation is prioritized, the use of application programming interfaces (APIs) to automate data exchange can be gradually introduced, and the migration of legacy systems to conform with data sharing protocols can be phased in over time. A data federation that is responsive to both needs and opportunities can be evolved through communities of practices that provide scientific guidance, by data owners and repository directors who coordinate efforts, by working groups that develop policies and promote best practices, and by funding agencies who provide funding for community processes, targeted investments in capabilities development and operations, and for adoption of standards in scientific workflows.

Opportunities and approaches

The exposomics field can take advantage of previous and ongoing investments from prior and existing data sharing efforts. For instance, the NIH has recently invested in a common single sign-on/multifactor authentication service, the Researcher Auth Service (RAS). The long-term benefits in enabling cross-site automation and of easing the access burdens that researchers confront daily through the adoption of technology like RAS arguably outweigh the short-term costs of adoption. Work under worldwide organizations promoting data sharing, such as the Research Data Alliance (RDA) and the Global Alliance for Genomics and Health (GA4GH) are developing protocols, toolkits, and reference implementation for needed capabilities like federated discovery, machine readable consent and usage forms, and data usage ontologies which the exposomics community can contribute to and adopt. Open-source data catalog and data sharing platforms (e.g., Molgenis, Gen3, DataVerse, OSF) can greatly lower costs of deploying new catalogs and repositories and foster greater standardization through easy sharing of platform data/metadata models. The recent development of secure data enclaves that support large-scale studies, such as All of Us and the National Covid Consortium (NC3), offer system architectures, design patterns, and best practices for deploying new enclaves and in cases specific technologies such as for privacy preserving record linkage.[35] The NIH funded Cloud Interoperability Program (NCIP), which is focused on interoperability between data resources,

provides methods, use cases, and solutions for the exposomics community to build upon.

In the areas of common language and standards, traditional medical and clinical standards such as the FHIR, Observational Medical Outcomes Partnership (OMOP) Common Data Model, and the Patient-Centered Outcomes Research Common Data Model (PCOR-CDM) are already expanding to include more social and environmental determinants of health. The Organization for Economic Co-operation and Development (OECD) has developed harmonized templates that exist in many areas of relevance. Efforts like the European Human Exposome Network (EHEN) metadata working group, the Environmental Health Language Collaborative (EHLC), GO-FAIR, PhenX, NIH Common Data Elements portal, and the NIEHS Disaster Research Response Program (DR2) are disseminating and promoting standards and working to address gaps using tools and frameworks such as CEDAR Workbench[36] and ISA-TAB.[37] These efforts can be readily leveraged to support the needs of the exposomics community.

Groups within the EHEN are developing technologies of interest for an exposomics data federation, including a metadata catalog built on the Molgenis platform[38] and the DataSHIELD platform[39] which enables analysis across federated data resources without moving the data from the secure environment of the data provider. This last capability provides a pathway towards broader privacy preserving federated analysis as demonstrated in recent work.[40] Industry interest in and use of privacy preserving federated learning and artificial intelligence (AI) has sparked research and development in this field (see Torkzadehmahani et al.[41] for recent review) that the exposomics field should seek to adopt into practice.

Access to sensitive data is still hampered by a diversity of data and the existing material transfer agreements, data use agreements, and subject consent language around data sharing. There are also needs for developing and adopting templates and the use of forms that are machine readable and harmonizable. While progress in these areas has been slow, the efforts like the International HundredK+ Cohorts Consortium (IHCC) and the GA4GH are seeking to advanced data access procedures that apply across studies at an international level and that should be applicable to exposomics studies. For instance, GA4GH has produced a toolkit for drafting machine-readable consent forms and a data access committee review standards toolkit to promote consistent review criteria and is actively working on improving access under its Regulatory and Ethics Work Stream.

Research and development of natural language models and their application for complex tasks such as text summarization, question answering, and information extraction has undergone tremendous steps forward in recent years, with large language model technologies transitioning to mainstream use. The approaches can be applied to federated systems to improve natural language search, translation of human language, translation of language across scientific subdomains, provide summarization of resources managed across federated systems, as well as to aid in producing informative visualizations such as evidence maps.

Recommendations

The following recommendations can be grouped into two broad categories. The items 1, 2, 3, 6, 7, and 10 focus on community and coordination activities that will require leadership and support from funders worldwide to foster as well as engagement from scientists, data stewards, directors of data repositories, and data scientists/informaticists to align and coordinate efforts. The items 4, 5, 8, and 9 focus on the piloting, development, upgrading, and maintenance of cyberinfrastructure to support the management, sharing, and usage of data sets. Existing funding call, such as the NIH funding for established Data Repositories and Knowledgebases, will help to support these efforts although it is likely that an additional support will be needed. Achieving these ambitious goals also calls for coordinating the funding across nations and individual federal agencies to maximize the return on investment and reduce duplication.

1. *Establish an exposome engineering task force.* The Internet Engineering Task Force (IETF) successfully operated a community-driven international model for evolving Internet protocols despite competing interests of participants. Adopting successful models such as IETF will help with evolving a federation that is useful for the exposome community and aligning efforts of organizations and stakeholders worldwide. This effort should align with organizations like GA4GH and RDA that are advancing interoperability protocols and technologies.
2. *Establish an exposome Community of Practice (CoP).* In addition to a task force that can recommend protocols and technical approaches, a CoP would provide an important communications channel between data repositories, providers of data services, tool builders, and users to communicate with and share and promote approaches and align efforts. Such communications are currently challenging as technical approaches are typically not documented in literature and are instead often communicated by word-of-mouth, which leads to silo'ing of approaches by communities and based on physical distance. An Exposome CoP will be well positioned to promote and advance a roadmap for federation, to topropose and promote driving use cases, and organize workgroups as useful to address topical topics.
3. *Develop and maintain a library of research use cases.* To fully guide investments and priorities in federation, the development and sharing of detailed scientific use cases with needed data (and data levels) is needed. The development of detailed use cases could be accomplished through break-out or working group sessions as part of large workshops and conferences with sessions arranged by the CoP to ensure researchers, data scientists, tool developers and repository representatives are present.
4. *Promote the adoption of core federation technologies among repositories.* Data repositories, data systems, and tools today will need some amount of new engineering to support features needed for federation, such as new application programming interfaces (APIs), curation of data to support standards needed for federation, adoption of enabling technologies like the NIH Researcher Auth Service (RAS) and DataShield. Targeted and supplemental funding will be needed to ensure capabilities for federation can be developed.
5. *Invest in piloting and adoption of approaches to mitigate privacy and data restriction issues.* Several approaches exist or have been proposed to address issues that restrict access to human data, such as creating trusted third-party hosting sites, use of machine-readable agreements to facilitate aspects of the data access process, use of federated analysis capabilities, use of secure multiparty computations across different repositories, and generating computable-encrypted data sets using homomorphic encryption. The

exposome community should continue to discuss, pilot, and adopt these approaches based on utility, coordinating through groups such as the proposed Engineering Task Force and CoP.

6. *Establish interoperability and competency challenges:* Challenges are often used to assess and ensure technology capabilities. This is especially important for federated systems given that coordination is required between distributed resources. Challenges also represent a way to assess the success of targeted funding. Challenges can start around limited use cases, such as the GA4GH notion of beacons ('identify all cohorts that have collected occupational history and spirometry measures') and evolve to more complex challenges.
7. *Foster the development and adoption of cross-repository data and metadata standards*. Efforts exist across communities to develop relevant standards, such as the EHEN metadata working group, clinical standards working groups (e.g., for FHIR, OMOP), and the Environmental Health Language Collaborative (EHLC). Currently, these efforts are voluntary with little recognition or funding support. Increasing the funding, devoted time, and recognition for researchers involved in such work will ensure this critical need is adequately addressed.
8. *Provide for sustained management and sharing of exposomics data*. All relevant exposomics data should be captured and maintained in a set of federated data repositories that are funded to be sustainable and able to meet FAIR and TRUST[42] principles and conform to the Desirable Characteristics for All Data Repositories (NIH NOT-OD-21-016).
9. *Provide a data analysis platform and establish and maintain a library of data analysis tools*. Data analysis platforms will allow the exposome researchers to conduct data analysis in the cloud without having to download large amounts of data onto their local machines or acquire and maintain expensive hardware. This practice will also protect the data from being spread to many different machines, ensuring the security and privacy of the data. One example of such a data analysis platform is the Jupyter Notebook that the All of Us program has adopted that enables researchers to directly write Python or R computer scripts for data analysis. The federated data ecosystem can establish and maintain a library of data analysis tools that the community has developed for the Jupyter Notebook environment so that other researchers do not have to reinvent the wheel. Accompanying this effort should be a well-written list and description of such tools for the community to use.
10. *Organize regular workshops and office hours*. It will be important to communicate with the exposure science community about the federated data ecosystem and educate the community about how to take advantage of the data resources and data analysis tools that the ecosystem provides.

Conclusion

Feedback from the 2023 exposomics workshop indicated a clear community need for a federated exposomics data ecosystem. While non-trivial challenges exist in creating this ecosystem, new data sharing policies and work to build data sharing ecosystems within the life sciences and other scientific communities can provide a foundation for success and advance exposomics research, if there is adequate support to enable an exposomics community wide effort.

Acknowledgements

C.P.S acknowledges financial support of the National Institute of Environmental Health Sciences. J.R.G. acknowledge financial support of the European Union's Horizon 2020 research and innovation programme under grant agreement No 874583 (ATHLETE project). A.R. was funded by NIH: NIDA grant R01 DA053028 "CRCNS: NeuroBridge: Connecting big data for reproducible clinical neuroscience." The authors would like to thank Jeremy Erickson, Jennifer Fostel, and David Fargo for insightful comments and suggestions.

Data availability

No new data were generated or analyzed in support of this research.

Conflict of interest statement

None declared.

Author contributions

Charles P. Schmitt (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Jeanette A. Stingone (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Arcot Rajasekar (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Yuxia Cui (Conceptualization [equal]), Xiuxia Du (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Chris Duncan (Conceptualization [equal]), Michelle Heacock (Conceptualization [equal]), Hui Hu (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Juan R. Gonzalez (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Paul D. Juarez (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Alex I. Smirnov (Conceptualization [equal], Writing—original draft [equal], Writing—review & editing [equal])

References

1. Harker J. Progress in exposomics, precision health made at scientific summit. October 2022. accessed August 30, 2023. https://factor.niehs.nih.gov/2022/10/science-highlights/exposome-summit
2. Manrai AK, Cui Y, Bushel PR, *et al.* Informatics and data analytics to support exposome-based discovery for public health. Annu Rev Public Health. 2017;38:279-294. https://doi.org/10.1146/annurev-publhealth-082516-012737
3. Turner MC, Nieuwenhuijsen M, Anderson K, *et al.* Assessing the exposome with external measures: commentary on the state of the science and research recommendations. Annu Rev Public Health. 2017;38:215-239. https://doi.org/10.1146/annurev-publhealth-082516-012802
4. Turner MC, Vineis P, Seleiro E, *et al.* EXPOsOMICS: final policy workshop and stakeholder consultation. BMC Public Health. 2018;18(1):260.https://doi.org/10.1186/s12889-018-5160-z

5. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.* The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018.https://doi.org/10.1038/sdata.2016.18
6. HDC HHEAR Data Center. Accessed August 30, 2023. https://hhearprogram.org/data-center
7. EE Exposome Explorer. Accessed August 30, 2023. http://exposome-explorer.iarc.fr/
8. Behrens T, Ge C, Vermeulen R, *et al.* Occupational exposure to nickel and hexavalent chromium and the risk of lung cancer in a pooled analysis of case-control studies (SYNERGY). Int J Cancer. 2023;152(4):645-660. https://doi.org/10.1002/ijc.34272
9. Hvidtfeldt UA, Chen J, Rodopoulou S, *et al.* Breast cancer incidence in relation to long-term low-level exposure to air pollution in the ELAPSE pooled cohort. Cancer Epidemiol Biomarkers Prev. 2023;32(1):105-113. https://doi.org/10.1158/1055-9965.Epi-22-0720
10. Hinchliffe A, Alguacil J, Bijoux W, *et al.* Occupational heat exposure and prostate cancer risk: a pooled analysis of case-control studies. Environ Res. 2023;216(Pt 2):114592.https://doi.org/10.1016/j.envres.2022.114592
11. Olsson A, Guha N, Bouaoun L, *et al.* Occupational exposure to polycyclic aromatic hydrocarbons and lung cancer risk: results from a pooled analysis of case-control studies (SYNERGY). Cancer Epidemiol Biomarkers Prev. 2022;31(7):1433-1441. https://doi.org/10.1158/1055-9965.Epi-21-1428
12. Yang JJ, Yu D, White E, *et al.* Prediagnosis leisure-time physical activity and lung cancer survival: a pooled analysis of 11 cohorts. JNCI Cancer Spectr 2022;6(2):pkac009. https://doi.org/10.1093/jncics/pkac009
13. Stingone JA, Buck Louis GM, Nakayama SF, *et al.* Toward greater implementation of the exposome research paradigm within environmental epidemiology. Annu Rev Public Health. 2017;38:315-327. https://doi.org/10.1146/annurev-publhealth-082516-012750
14. Zhang Y, Hu H, Fokaidis V, *et al.* Identifying environmental risk factors for post-acute sequelae of SARS-CoV-2 infection: an EHR-based cohort study from the recover program. Environ Adv. 2023;11:100352.https://doi.org/10.1016/j.envadv.2023.100352
15. Global Burden of Disease Long COVID Collaborators. Estimated Global Proportions of Individuals With Persistent Fatigue, Cognitive, and Respiratory Symptom Clusters Following Symptomatic COVID-19 in 2020 and 2021. *JAMA*. 2022;328(16):1604-1615. https://doi.org/10.1001/jama.2022.18931
16. Weber GM, Murphy SN, McMurry AJ, *et al.* The shared health research information network (SHRINE): a prototype federated query tool for clinical data repositories. J Am Med Inform Assoc. 2009;16(5):624-630. https://doi.org/10.1197/jamia.M3191
17. Visweswaran S, Becich MJ, D'Itri VS, *et al.* Accrual to clinical trials (ACT): a clinical and translational science award consortium network. JAMIA Open. 2018;1(2):147-152. https://doi.org/10.1093/jamiaopen/ooy033
18. Maitre L, Guimbaud JB, Warembourg C, *et al.* State-of-the-art methods for exposure-health studies: results from the exposome data challenge event. Environ Int. 2022;168:107422. https://doi.org/10.1016/j.envint.2022.107422
19. Johnson CH, Athersuch TJ, Collman GW, *et al.* Yale school of public health symposium on lifetime exposures and human health: the exposome; summary and future reflections. Hum Genomics. 2017;11(1):32.https://doi.org/10.1186/s40246-017-0128-0
20. Hartung T. A call for a human exposome project. Altex 2023;40(1):4-33. https://doi.org/10.14573/altex.2301061
21. Sillé FCM, Karakitsios S, Kleensang A, *et al.* The exposome—a new approach for risk assessment. Altex 2020;37(1):3-23. https://doi.org/10.14573/altex.2001051
22. Vermeulen R, Schymanski EL, Barabási A-L, Miller GW. The exposome and health: where chemistry meets biology. Science. 2020;367(6476):392-396. https://doi.org/10.1126/science.aay3164
23. Stathias V, Turner J, Koleti A, *et al.* LINCS data portal 2.0: next generation access point for perturbation-response signatures. Nucleic Acids Res. 2020;48(D1):D431-D439. https://doi.org/10.1093/nar/gkz1023
24. McMahan HB, Moore E, Ramage D, Hampson S, Agüera y Arcas B. (2016). *Communication-efficient learning of deep networks from decentralized data*. arXiv. 2023. https://doi.org/10.48550/arXiv.1602.05629
25. sharing.nih.gov. NIHGDSP, data sharing and release xpectations. Accessed August 30, 2023. https://sharing.nih.gov/genomic-data-sharing-policy/submitting-genomic-data/data-submission-and-release-expectations
26. Bohn R, Lee C, Michel M. *The NIST Cloud Federation Reference Architecture, Special Publication (NIST SP)*. National Institute of Standards and Technology; 2020. https://doi.org/10.6028/NIST.SP.500-332
27. Brown TB, Mann B, Ryder N, *et al.* Language Models are Few-Shot Learnings. *arXiv*. 2020. https://doi.org/10.48550/arXiv.2005.14165
28. Juarez PD, Hood DB, Song MA, Ramesh A. Use of an exposome approach to understand the effects of exposures from the natural, built, and social environments on cardio-vascular disease onset, progression, and outcomes. Front Public Health. 2020;8:379.https://doi.org/10.3389/fpubh.2020.00379
29. Martin-Sanchez F, Bellazzi R, Casella V, Dixon W, Lopez-Campos G, Peek N. Progress in characterizing the human exposome: a key step for precision medicine. Yearb Med Inform. 2020;29(1):115-120. https://doi.org/10.1055/s-0040-1701975
30. Zhang P, Carlsten C, Chaleckis R, *et al.* Defining the scope of exposome studies and research needs from a multidisciplinary perspective. Environ Sci Technol Lett. 2021;8(10):839-852. https://doi.org/10.1021/acs.estlett.1c00648
31. Barnes C, Bajracharya B, Cannalte M, *et al.* The biomedical research hub: a federated platform for patient research data. J Am Med Inform Assoc. 2022;29(4):619-625. https://doi.org/10.1093/jamia/ocab247
32. Chaterji S, Koo J, Li N, Meyer F, Grama A, Bagchi S. Federation in genomics pipelines: techniques and challenges. Brief Bioinform. 2019;20(1):235-244. https://doi.org/10.1093/bib/bbx102
33. Thorogood A, Rehm HL, Goodhand P, *et al.* International federation of genomic medicine databases using GA4GH standards. Cell Genom 2021;1(2):100032. https://doi.org/10.1016/j.xgen.2021.100032
34. Vrijheid M, Slama R, Robinson O, *et al.* The human early-life exposome (HELIX): project rationale and design. Environ Health Perspect. 2014;122(6):535-544. https://doi.org/10.1289/ehp.1307204
35. Kiernan D, Carton T, Toh S, *et al.* Establishing a framework for privacy-preserving record linkage among electronic health record and administrative claims databases within PCORnet, the national patient-centered clinical research network. BMC Res Notes. 2022;15(1):337.https://doi.org/10.1186/s13104-022-06243-5
36. Musen MA, O'Connor MJ, Schultes E, Martínez-Romero M, Hardi J, Graybeal J. Modeling community standards for metadata as templates makes data FAIR. Sci Data. 2022;9(1):696.https://doi.org/10.1038/s41597-022-01815-3
37. Rocca-Serra P, Sansone S-A, Brandizi M. Specification documentation: ISA-TAB 1.0. Zenodo; 2009. https://doi.org/10.5281/zenodo.161355

38. van der Velde KJ, Imhann F, Charbon B, *et al.* MOLGENIS research: advanced bioinformatics data software for non-bioinformaticians. Bioinformatics. 2019;35(6):1076-1078. https://doi.org/10.1093/bioinformatics/bty742
39. Wolfson M, Wallace SE, Masca N, *et al.* DataSHIELD: resolving a conflict in contemporary bioscience—performing a pooled analysis of individual-level data without sharing the data. Int J Epidemiol. 2010;39(5):1372-1382. https://doi.org/10.1093/ije/dyq111
40. Escribà-Montagut X, Marcon Y, Avraam D, *et al.* Software application profile: ShinyDataSHIELD—an R Shiny application to perform federated non-disclosive data analysis in multicohort studies. Int J Epidemiol. 2023;52(1):315-320. doi:10.1093/ije/dyac201
41. Torkzadehmahani R, Nasirigerdeh R, Blumenthal DB, *et al.* Privacy-preserving artificial intelligence techniques in biomedicine. Methods Inf Med. 2022;61(S 01):e12-e27. https://doi.org/10.1055/s-0041-1740630
42. Lin D, Crabtree J, Dillo I, *et al.* The TRUST Principles for digital repositories. Sci Data. 2020;7(1):144. https://doi.org/10.1038/s41597-020-0486-7