National Advisory Dental and Craniofacial Research Council
Data Science Strategy Working Group

Final Report

May 2024

NIH
National Institute of Dental and Craniofacial Research

Table of Contents

Executive Summary

Data science and the opportunities arising from it have become transformative for the biomedical sciences. This potential has been acknowledged by the National Institutes of Health (NIH) and led to the development of an NIH-wide data science strategy. Concurrently, the increasing importance of data science in dental, oral, and craniofacial (DOC) research has been recognized by the National Institute of Dental and Craniofacial Research (NIDCR). The current NIDCR strategic plan identifies multiple challenges and opportunities related to data science across the translational spectrum of DOC research, clinical care, and community support.

To address these challenges and opportunities, the NIDCR Director, Dr. Rena D'Souza, assembled and charged the National Advisory Dental and Craniofacial Research Council (NADCRC) Data Science Strategy Working Group (DSS-WG) with providing recommendations on developing a data science strategy that would set the course for future implementation. The DSS-WG consisted of members covering the entire translational spectrum of DOC research. The DSS-WG held regular meetings and conducted targeted information gathering activities from September 2022 to January 2024. The present report summarizes the findings and recommendations of the working group.

In the information gathering phase, the DSS-WG investigated the state of the current DOC data ecosystem, including its structure, challenges, and opportunities. To complement the expertise of working group members, the DSS-WG conducted listening sessions with DOC community members and worked with NIDCR staff to conduct a request for information (RFI). These information gathering activities resulted in an extensive catalog of data types, data standards, data sources, and data systems currently used by the DOC community. These activities also identified technical, scientific, academic, regulatory, and resource-related challenges to data sharing.

The DSS-WG observed that the current DOC data ecosystem is characterized by substantial complexity and heterogeneity, with large numbers of data types and standards being used across dozens of data systems. Importantly, the DOC data ecosystem lacks well-defined limits, encompassing both DOC-specific resources and generalist data systems that are widely used by DOC researchers. While some existing systems are connected, in many cases the lack of interoperability between systems can reduce findability of data and constitutes a hurdle to integrative analysis. In some cases, funding mechanisms that are designed for research projects of limited duration may interfere with long-term sustainability of data systems, impeding progress toward compliance with FAIR (findable, accessible, interoperable, reusable) data principles.

The DSS-WG identified several areas in which the specific characteristics of DOC research, clinical care, and community needs create opportunities in data science that are unique to NIDCR. Addressing these opportunities requires alignment with the overarching NIH data science strategy. The DSS-WG proposes that integrating the diversity of DOC data, including genomic, imaging, and population health data, provides unprecedented avenues for understanding and addressing complex oral health issues. This includes the potential for new insights into the genetic, environmental, and social factors influencing oral health disparities. The integration of advanced artificial intelligence (AI) and machine learning (ML) techniques with the diverse DOC

data sets has potential for enhanced diagnostic and treatment strategies and offers more personalized and effective oral healthcare solutions. Emphasizing the necessity of data integration, the DSS-WG underscores the value of unified strategy, governance, and standards for data management, sharing, and use within DOC research. This unified approach would aim to coordinate and promote consistent handling of data across different research domains and projects in the DOC field, ensuring that data is managed, shared, and used in ways that optimize research advancement, enhance patient outcomes, and more effectively meet community health needs.

This report provides recommendations toward the development of an NIDCR data science strategy. These recommendations are organized within the framework of the NIH Strategic Plan for Data Science[1] and address the specific challenges of DOC research. The recommendations include:

1. NIDCR should establish a robust data infrastructure that is specifically tailored for DOC research, while interfacing with relevant NIH data systems.
2. NIDCR should establish new and modernize existing DOC-specific data ecosystems.
3. NIDCR should foster the development of data management, analytics, and visualization tools for DOC research.
4. NIDCR should enhance workforce development in data science, both internally and across the research community. This should include inclusive training programs to promote the diversity of the future DOC data science research workforce.
5. NIDCR should promote data stewardship and sustainable data policies to ensure the integrity, confidentiality, and FAIRness of data in DOC research.

The DSS-WG conducted a detailed assessment of the DOC data ecosystem and its resulting recommendations lay a solid foundation for a well-informed NIDCR data science strategy, enabling researchers to realize the potential of data science across the full translational spectrum of DOC research, addressing health disparities, and improving overall health.

[1] NIH Strategic Plan for Data Science, available at https://datascience.nih.gov/strategicplan

Towards an NIDCR Data Science Strategy

Alignment with NIDCR Strategic Priorities

The *NIH-Wide Strategic Plan* released in 2021[2] clearly articulates NIH's commitment to building data resources that can support research progress. Such foundational resources will enable and advance basic research and accelerate the understanding of the biological and environmental factors that contribute to human health and disease. Importantly, the development of such resources must be guided by the FAIR data principles (findable, accessible, interoperable, reusable)[3] to maximize the impact of NIH-funded research by enabling a multitude of downstream and integrative analyses on any given data set produced.

To provide a roadmap for the future of the biomedical data science ecosystem, in 2018 the NIH Office of Data Science Strategy released the first *NIH Strategic Plan for Data Science*[4]. Defining data science as "the interdisciplinary field of inquiry in which quantitative and analytical approaches, processes, and systems are developed and used to extract knowledge and insights from increasingly large and/or complex sets of data", this plan expands on the opportunities for the advancement of biomedical research that arise from data science. The plan also outlines initial sets of general challenges associated with biomedical data science, including the growing cost of managing data, lack of integration across specialized data resources, lack of standardization, and a lack of persistence due to funding structures. The plan describes an ambitious set of goals and objectives to be implemented across NIH, including the creation and integration of data infrastructure, modernization of the data ecosystem, the development of data management and analysis tools, the development of the NIH data science workforce, and data stewardship. All of these considerations apply in general terms to the DOC data ecosystem, including those resources under the purview of NIDCR, but are intertwined with multiple considerations that are specific to the DOC data ecosystem, as outlined in more detail throughout this report.

The current *NIDCR Strategic Plan*[5], released in 2021, describes five strategic priorities. Importantly, all five priorities will require advancements of data science and the DOC data ecosystem in order to succeed. Selected examples of specific objectives that will require such advancements to be successful are shown in **Table 1**.

Considering the critical need of data and data science in order to achieve success in all of NIDCR's strategic priorities, it is necessary to develop a more detailed roadmap to guide these efforts in the years ahead. Importantly, such a roadmap needs to be developed in the context of overarching NIH priorities, but with the unique challenges and considerations of DOC research and the existing DOC data ecosystem in mind.

---

[2] NIH-Wide Strategic Plan Fiscal Years 2021-2025, available at https://www.nih.gov/about-nih/nih-wide-strategic-plan

[3] FAIR Principles, available at https://www.go-fair.org/fair-principles/

[4] NIH Strategic Plan for Data Science, available at https://datascience.nih.gov/strategicplan

[5] NIH Publication No. 22-DR-8175, available at https://www.nidcr.nih.gov/sites/default/files/2022-01/NIDCR-Strategic-Plan-2021-2026.pdf

NIDCR Strategic Priorities 2021-2026	Examples of Data-Related Objectives
Priority #1: Integrate Oral and General Health *Advance discoveries across the translational research spectrum and drive innovations that improve the early diagnosis, prevention, and treatment of DOC diseases across the life span.*	• **Integration** of DOC conditions with systemic diseases using *All of Us* Research Program data. • Unique and shared **taxonomies** for DOC research.
Priority #2: Precision Dental Medicine *Develop more precise and individualized treatments for the management and prevention of DOC diseases.*	• **Expand data ecosystems** that use computational tools and mobile technologies to improve health outcomes in individuals and specific populations. • Rapid and less costly point-of-care technologies with **high resolution data** capture that improve patient health outcomes and facilitate remote or virtual tele-dentistry.
Priority #3: Translate and Implement *Accelerate the translation of research and the implementation of new discoveries into oral and general healthcare practices that reduce health inequities and disparities and improve oral health outcomes for individuals and communities worldwide.*	• Novel platforms and advanced technologies that gather medical, dental health, personal, and other health systems data into **electronic health records**. • Standardized DOC disease-related **ontologies**.
Priority #4: Diverse Research Pipeline *Nurture future generations of DOC researchers and oral health professional scholars who can address public health needs within a continually evolving landscape of science and technology advances.*	• Create new training and career development programs that engage and recruit students and postdoctoral researchers to **harness the power of data science** applications.
Priority #5: Partner and Collaborate *Expand existing partnerships and create new ones to advance the NIDCR research enterprise and increase its reach and impact.*	• **Knowledge-sharing** with stakeholder groups through outreach, and identification of gaps in knowledge that relate to public health challenges.

Table 1: NIDCR Strategic Priorities and Data-Related Objectives.

The NIDCR Data Science Strategy Working Group

In the fall of 2022, the NADCRC Data Science Strategy Working Group (DSS-WG) was assembled to provide recommendations on developing a data science strategy to complement and support the NIDCR Strategic Plan. The working group was not asked to develop a strategic plan, per se, but was tasked with assessing the current state of the DOC data ecosystem and challenges associated with it, synthesizing these findings, and making general recommendations that can inform and guide the development of an NIDCR data science strategy.

The DSS-WG was asked to consider the following questions:

- **How can the DOC data ecosystem evolve to facilitate research, research training, and research career development across the full translational spectrum of basic, preclinical, clinical, implementation, and public health research?**
- **How can DOC researchers apply data and data science methods to expedite the development and delivery of oral health solutions to all individuals?**
- **How can we power research in health disparities and inequalities with data and data science methods to inform strategies for overcoming the disparities and inequalities?**

The group was asked to address these questions covering the entire translational spectrum of DOC research, including T0 (basic biomedical research), T1 (translation to humans), T2 (translation to patients), T3 (translation to practice), T4 (translation to communities). Consequently, the twelve members of the DSS-WG (see **Appendix 1**) were selected to represent the entire translational spectrum.

From September 2022 to January 2024, the Working Group held regular meetings and conducted additional activities, such as community listening sessions, to gather and analyze information, synthesize findings and conclusions, and develop recommendations to guide the development of an NIDCR Data Science Strategy. Findings and recommendations were presented in regular updates by the Working Group Chair to the NADCRC and are summarized in the present report.

Introducing the Current Landscape of Data Sharing

To provide general context and define the terminology used throughout this report, this section gives an overview of current approaches to data sharing across clinical and life science domains. We begin by defining the specific terms used in this report, categorize the general types of repositories, and outline the technical and social context of data use. Finally, we highlight several community leadership efforts relevant to the topic of this report.

Definitions of Terms

Table 2 defines the terms relevant to data and data ecosystems as they are used in this report.

Term	Definition
data	Measurements, quantities, or qualities that are represented in a form suitable for analysis.
data ecosystem	A network of actors (enterprises, institutions, individuals) and data-related resources (data sets, software, infrastructure) in which the actors carry out various duties and activities to produce, curate, manage, share, and make use of data sets[6,7].
data infrastructure	The computational hardware and software to support data storage, access, and analysis, as well as the organizational knowledge and policies to support these tasks.
data set	A structured collection of data that is the product of an experiment, observational study, or analysis.
data repository	A collection of data sets that is organized and managed to facilitate archiving, retrieval, and sharing of the data sets.
document	A collection of information that is arranged in a specified order and intended to inform about some specific topic. Documents usually rely on text to convey information, and may include figures, data tables, or small data sets.
knowledgebase	A repository of curated information that represents a shared understanding within a domain of inquiry.
metadata	Data that describes a data set. Examples are the creator, standards used in data representation, method of collection, and copyright.
research product	Data or information created for the purpose of documenting or communicating research.

Table 2: Data and Data Ecosystem Terminology.

[6] Oliveira et al., *Investigations into Data Ecosystems: a systematic mapping study* (2019). Knowledge and Information Systems 61:589-630; available at https://doi.org/10.1007/s10115-018-1323-6

[7] Oliveira et al., *What is a data ecosystem?* (2018). Proceedings of the 19th Annual International Conference on Digital Government Research 74:1-9; available at https://doi.org/10.1145/3209281.3209335

Types of Repositories

Different types of data and information are accommodated by different types of repositories. **Table 3** categorizes different types of repositories and provides examples of each.

Type	Examples
Generalist repositories for sharing research products[8]	Dataverse Dryad Figshare Mendeley Data Vivli Zenodo *each supported by the Generalist Repository Ecosystem Initiative*
Domain-specific repositories for sharing specific types of data or data specific to a discipline[9]	The NIH website on Scientific Data Sharing[10] provides a list of NIH-supported data repositories[11]. Some accept data from a broad set of investigators, while others maintain data only for a particular project.
Repositories that serve as platforms for supporting **collaborative work** and data sharing	Open Science Framework[12] *supported by the Generalist Repository Ecosystem Initiative*

Table 3: Types of Repositories.

[8] Generalist Repositories available at https://www.nlm.nih.gov/NIHbmic/generalist_repositories.html
[9] Domain-Specific Repositories available at https://www.nlm.nih.gov/NIHbmic/domain_specific_repositories.html
[10] Scientific Data Sharing available at https://sharing.nih.gov/
[11] Repositories for Sharing Scientific Data available at https://sharing.nih.gov/data-management-and-sharing-policy/sharing-scientific-data/repositories-for-sharing-scientific-data
[12] Open Science Framework available at https://osf.io/

Technical Context: Levels of Data Interoperability

Data ecosystems require data interoperability, which requires standards and coordination at four levels described with examples in **Table 4**.

Level	Explanation	Examples
semantic	The meaning and context of the data expressed through defined terms in ontologies and terminologies	Human Phenotype Ontology (HPO) SNOMED-CT Uberon ICD-O NCI Thesaurus
syntactic	The language expressing the data through data models, data structures, data dictionaries, and data schemes	OMOP Common Data Model FHIR BRIDG LinkML Bioschemas MIAME
system	The presentation of the data through common formats for representing, encoding, and decoding the data	OWL RDF VCF FASTA PFB
structural	The architecture of networks, applications, and web services	Docker various application programming interfaces (APIs)

Table 4: Levels of Data Interoperability. Content adapted from Melissa Haendel's presentation at the 2021 FaceBase Community Forum

Social Context: Using Data to Improve Health

Data ecosystems are formed by interactions among individuals, organizations, data resources, network infrastructure, and software platforms. The term *ecosystem* implies that all actors benefit from these interactions.

Data ecosystems may be designed for only a small number of organizations and individuals, or they may have many contributors and users. Some are "open", meaning the data is publicly available, and others are "closed" without outside access. As a special case, some systems are generally publicly available, but require registration or the data they contain has use restrictions (e.g., related to protection of patient data). **Table 5** lists questions to consider when creating and using data ecosystems.

Role	Relevant questions
Production of data *by researchers*	• What is the responsibility of data producers to annotate data using community standards? • How are the methods and processes used to collect the data documented? • Do researchers have the responsibility to use open file formats, rather than proprietary formats? • For patient data, how are issues of privacy and consent addressed? • Will the data and metadata support reproducible research?
Curation of data *by repository teams*	• Are there repository standards for curating the data and metadata so that users can search and browse across the repository?
Stewardship of data and management of repository *by repository teams*	• How are repositories funded to ensure sustainable, long-term access to data? • As standards evolve, will older data need to have new annotations or file formats? • Is data "AI-ready"?
Development of platforms *(tools and services) for integrating and searching data across repositories*	• Which repositories are accessed by the platform to build the ecosystem? • Whose needs are addressed by the platforms? • Are decisions about standards governed at the level of repositories or platforms?
Use of data *for discovery and decision making*	• What value is provided by the data? Who benefits? • How does a repository ensure data is trustworthy?

Table 5: Considerations When Creating or Using Data Ecosystems.

Additional factors that affect data sharing and reuse within data ecosystems include:

- Regulations and policies enacted by organizations and governmental bodies.
- Financial and economic policies that promote or discourage data reuse.
- The availability of training to build researchers' skills and knowledge relevant to data sharing and reuse.

Community Leadership for Data Sharing and Reuse

A number of organizations work to foster data sharing and reuse by establishing standards, helping researchers to locate repositories and data, and fostering discussions around the sharing and reuse of data:

- DataCite – non-profit that provides persistent identifiers (DOIs) for research data and other research products
- re3data – global registry of research data repositories
- FAIRsharing.org – organization that provides both leadership in FAIR-enabling activities and a curated repository of standards, databases, and policies
- FAIRsFAIR – Europe-based project to implement FAIR principles in data-sharing infrastructure
- OBO Foundry – community of developers of biomedical ontologies
- FORCE11 – community that seeks to transform scholarly communication
- Future of Privacy Forum – non-profit that provides leadership on issues of privacy protections, ethical norms, and business practices in response to challenges posed by technological innovation
- Research Data Alliance (RDA) – community-driven international initiative supporting the goal of building social and technical infrastructure to enable open sharing and re-use of data
- Data Curation Network (DCN) – membership organization of institutional and non-profit data repositories aimed at advancing open research by making data more ethical, reusable, and understandable

NIH Initiatives for Data Sharing and Data Ecosystems

The Office of Data Science Strategy works to coordinate FAIR data practices across NIH activities, including biomedical data repositories and knowledgebases.

NIH currently supports a number of efforts for data sharing and data ecosystems, including:

- The Common Fund Data Ecosystem provides a portal for searching the Common Fund data sets.
- The HEAL data ecosystem was created to address the opioid public health crisis and the NIH HEAL Data resources offer data and resources for data use. NIDCR is a key contributor to this initiative.
- The Genomic Data Commons of NCI provides a data portal for sharing and accessing cancer genomic studies
- The vision for the NCI Cancer Research Data Ecosystem is to advance precision medicine for cancer care by promoting sharing of data among all stakeholders.
- The *All of Us* Research Program collects health data from a diverse group of participants from across the United States. It provides a browser for aggregate-level data as well as a researcher workbench.

- The National COVID Cohort Collaborative (N3C) Data Enclave is a secure platform for harmonized clinical data to which NIDCR and DOC community made significant contributions.

With the NIH Data Management and Sharing Policy, which took effect in January 2023, funded investigators and institutions are being held to higher standards that will promote data sharing.

A Focus on DOC Data

Collecting Community Input

The DSS-WG performed an initial internal discovery exercise aimed at collating data resources, data types, and data standards by members across the translational spectrum in their respective work. Recognizing the still limited spectrum of activities of working group members compared to the diverse and multifaceted work of the broader DOC community, the working group additionally sought to engage a larger number of group-external researchers to collect input on data use and associated challenges across the entire translational continuum. For this purpose, the DSS-WG used two primary mechanisms for discovery: Listening sessions and a Request for Information (RFI).

Listening Sessions

In July 2023, the working group conducted two listening sessions with invited community participants who conduct data science and data intensive research in DOC biomedicine spanning the translational spectrum. One session covered T0-T2, a second session covered T3-T4 (see section **The NIDCR Data Science Strategy Working Group** above for definitions of T-levels).

In each listening session, participants shared their knowledge and experience addressing the following questions:

- *Questions for data generators:*
    - What kinds of data does your lab generate?
    - How do you currently store your data?
- *Questions for data users:*
    - What kinds of data does your lab use?
    - What is the source of the data that you use for your research?
- *Questions for participants who share DOC-specific data sets their team generates:*
    - How often do you share data with collaborators or make it publicly available?
    - How do you share data?
    - What barriers to data sharing do you experience?
- *Questions related to data standards:*
    - Do you use any common data standards, data elements, ontologies etc. in your research?
    - If yes, which of them are most important for your work?
    - If no, why not?
- *Open-ended question:*
    - What features do you consider important for a DOC data repository?

Request for Information (RFI)

Complementary to organizing Listening Sessions, the Working Group worked with NIDCR to issue a Request for Information (RFI) titled *"National Institute of Dental and Craniofacial Research (NIDCR) Strategic Planning of Infrastructure and Resources for Data Science Research and*

Research Training", NOT-DE-23-008[13]. The RFI was released on June 27, 2023, and remained open for responses until September 25, 2023.

The RFI invited responses from stakeholders and experts across the full spectrum of DOC research, including individual research laboratories, staff from scientific instrumentation core facilities, offices of research or sponsored projects, offices of provost, libraries, information technology and security personnel, institutional review boards, bioinformaticians, and data scientists and data managers who assist in data curation, formatting, and analysis.

Questions in the RFI, which used a web survey form, mirrored those presented in the listening sessions, with some changes and additions:

- Please indicate which of the groups of investigators you represent (e.g., Biostatistics, Collection and processing of biospecimens, Dentist, Development of experimental methods, Diet and nutrition, Epidemiology, Etiology of DOC conditions, Health disparities, Health services research, Identification of biomarkers, Oral metabolism and the microbiome, Pharmacist, Physician, Public health, Quality assurance/quality control, Treatment effectiveness and efficacy, other)
- Do you generate DOC-specific data sets and/or data sets that contain some DOC data in your own lab or institution? If so, how do you currently store your data?
- Do you share DOC data sets that you or your team/institution have generated? If so, how often and through what mechanism (e.g., encrypted email, shipped physical storage/thumb drives/hard disk drives, download/upload from website, cloud storage services, API/programmatic interface, file sharing services)?
- What challenges/barriers (if any) have prevented you or your lab from sharing data with the larger research community?
- Have you ever deposited DOC data in a repository? If so, what was your experience? If not, what barriers prevented you from doing so?
- Thinking about the last time you shared data, how much time did you or your staff spend preparing your data so it would be ready to share?
- How much of your research budget (%) is typically allocated to data management and sharing (e.g., repository fees, data manager salary)?
- Where do you go to find DOC data sets and resources that are relevant to your research?
- Which databases (data repositories or knowledgebases) have you used for your research?
- How are the databases and/or repositories that you use funded?
- Have you ever used shared or publicly available DOC data in your research? If so, what types of data do you usually use or look for?
- If you have retrieved publicly available DOC data in your research, how often do you find enough information supplied with the data (e.g., data provenance, data dictionary, metadata, etc.) to allow use?
- Have you integrated/merged data from multiple publicly available DOC data sets or integrated/merged them with some other type of healthcare data?
- If you have ever used shared or publicly available DOC data in your research, how would you characterize your overall experience of finding and using DOC data?

[13] NOT-DE-23-008, available at https://grants.nih.gov/grants/guide/notice-files/NOT-DE-23-008.html

- Do you currently use common data standards, data elements, ontology terms for describing your data, or common data models in your research? If so, which common data standards, elements, ontologies or models have you used and which of them are most important to your ability to utilize data in research? If not, what prevented you from doing so?
- Shared or publicly available data from which areas of the translational spectrum (T0-T4) are useful for your research purposes?
- In the course of your research/study, at what stage do you usually first make considerations for data management?
- In the course of your research/study at what stage do you usually first make considerations for data sharing?
- What software/tools are most useful to you in your current DOC research?
- What criteria do you use to select software and informatics tools for your work?
- Is the majority of your data analysis conducted locally or in the cloud?
- Thinking about the last time you shared data or software, did you provide any additional materials or information to help users understand and reuse your data or code?

Summary of Community Input Received

In total, we received input from 44 community members, including 18 participants in the T0-T2 listening session, 11 participants in the T3-T4 listening session, and 15 completed responses to the RFI. The following sections provide summaries of the input received from these discovery efforts, combining information collected in listening sessions and received through the RFI.

Data Types Generated or Used in DOC Research

Participants reported a large number of different data types that they generate or use in their research, which are summarized in **Table 6**. Despite the working group's broad outreach and discovery efforts, this list is likely incomplete. The list includes data types reported by participants in the listening sessions, by individual responses to the RFI, and by the DSS-WG members, who represent research areas across the translational spectrum, in an initial data gathering exercise. Data types are grouped by the type of entity described by the data.

Data Types	
Type of entity described by the data	**Data types**
Genetic, molecular, and biochemical characteristics of cells and tissues (both human and model organisms)	***Gene, gene expression, genomic, and other sequencing data*** • ATAC-seq (Assay for Transposase-Accessible Chromatin using sequencing) • ChIP-seq (chromatin immunoprecipitation sequencing) • DNA sequences by methods such as chain termination sequencing • functional genomics data from primary human tissues • genotyping

Data Types	
Type of entity described by the data	**Data types**
	• GWAS (genome-wide association studies), including summary statistics • QTL (quantitative trait locus analysis) • reporter gene assay • RNA in situ hybridization and transcript expression location detection by hybridization chain reaction • single-cell and single-nucleus ATAC-seq • single-cell and single-nucleus RNA-seq • SNP (single nucleotide polymorphism) • transcription profiles, such as microRNA profiles • WES (whole exome sequencing) • WGS (whole genome sequencing) ***Molecular and biochemical assay data*** • flow cytometry data • microbiological parameters (e.g., growth, pH, metabolism) • multiplex immunoassay data • metabolipidomics • metabolomics ***Microscopy imaging*** • fluorescence microscopy, including confocal microscopy • hematoxylin and eosin (H&E) stain • optical microscopy • laser capture microdissection (LCM) • Raman microscopy • SEM (scanning electron microscopy) • TEM (transmission electron microscopy) • atom probe tomography ***Spectroscopy*** • micro infrared spectroscopy • microbeam particle-induced X-ray emission spectroscopy
Microbial populations	• metagenomics data • microbial gene expression data • microbiome data
Anatomy of small structures and small animals	***Microscopy imaging*** • micro-CT (computed tomography) • optical projection tomography (OPT)
Disease, disorder, and treatment of patients	***Clinical encounter data*** (both structured and unstructured) • caries exam data • diagnostic data (coded as ICD) • treatment data (coded as CPT and HCPCS) • outcome data • medication data • cancer-related data, including expression data sets ***Clinical imaging data*** • CT and CBCT (cone beam computed tomography) images, including facial CTs

Data Types	
Type of entity described by the data	**Data types**
	• dental models • digital whole slides of diagnostic tissue • magnetic resonance imaging (MRI), including facial MRIs • micro-MRI • facial photographs • ultrasound images ***Diagnostic data streams*** • EEG (electroencephalogram) • Speech samples (audio and video) from patients with cleft and craniofacial anomalies ***Morphology descriptors and phenotype classifications*** • dental phenotypes • craniofacial phenotypes • morphometric analysis ***Clinical trial data*** • response to treatment • adverse event data • studies of craniofacial birth defects and orofacial cleft
Patient characteristics and experiences	***Patient characteristics (non-clinical)*** • demographics, including age, sex, race, ethnicity, and socioeconomic status • family history • insurance status and carrier • zip code or other geographic information ***Patient-reported experiences*** • patient-reported outcomes, including pain • patient-reported behaviors (may use questionnaires, surveys) • patient-reported risk factors
Perceptions and behavior of research participants	• survey and focus group data • survey data from communities and patients
Biospecimens	• data about biospecimens

Table 6: Data types used in DOC research.

Standards for Data and Systems Interoperability Used in DOC Research

In **Table 7**, we list terminologies and ontologies, data formats, and file formats used by DOC researchers. The list includes entries reported by participants in the listening sessions, provided through the RFI, and reported by DSS-WG. The relatively short list (compared to data types and data resources used by DOC researchers) may be related to the absence of suitable standards for many of the data types commonly used by DOC researchers, as described in more detail in the section **Challenges to Data Sharing** below.

Level of Interoperability (as defined in introduction)		**Used in DOC Research**
semantic	The meaning and context of the data expressed through defined terms in ontologies and terminologies	• CTCAE (Common Terminology Criteria for Adverse Events) • GO (Gene Ontology) • HPO (Human Phenotype Ontology) • ICD (International Statistical Classification of Diseases and Related Health Problems) codes • LOINC (Logical Observation Identifiers Names and Code) • MP Ontology (Mammalian Phenotype Ontology) • NCIt (National Cancer Institute Thesaurus) • SNODENT (Systematized Nomenclature of Dentistry) including SNODDS • SNOMED (Systematized Nomenclature Of Medicine)
syntactic	The language expressing the data through data models, data structures, data dictionaries, and data schemes	• caDSR (Cancer Data Standards Registry and Repository) • FHIR (Fast Healthcare Interoperability Resources) data standards for oncology • i2b2 (Informatics for Integrating Biology & the Bedside) common data model • JSON (JavaScript Object Notation) • mCODE (Minimal Common Data Elements) data standards for oncology • OMOP (Observational Medical Outcomes Partnership) • PCORnet (National Patient-Centered Clinical Research Network) Common Data Model • PhenX Toolkit (consensus measures for Phenotypes and eXposures) standards
system	The presentation of the data through common formats for representing, encoding, and decoding the data	• BAM (Binary Alignment Map) • FASTA • FASTQ • TSV (tab-separated value) format
structural	The architecture of networks, applications, and web services	• *none identified*

Table 7: Standards for Data and Systems Interoperability Used in DOC Research.

Data Resources and Repositories Used in DOC Research

Working group members, participants in listening sessions, and respondents to the RFI collectively reported nearly 60 different data systems, databases, data resources, and web sites which they contribute data to or which they use in their work as a source of data or information. As described below in more detail, this includes the full range from specialized, DOC-focused systems to general-purpose databases used by a wide range of disciplines.

Relevant data systems, sources, and repositories include:

- AAOF (American Association of Orthodontists Foundation) Legacy Collection: www.aaoflegacycollection.org
- ADA (American Dental Association) Masterfile
- *All of Us* Research Program: allofus.nih.gov
- BigMouth Dental Data Repository: bigmouth.uth.edu
- BRFSS (Behavioral Risk Factor Surveillance System): www.cdc.gov/brfss
- Bridge2AI (Bridge to Artificial Intelligence): bridge2ai.org
- BWHS (Black Women's Health Study): www.bu.edu/bwhs
- Census data
- CHIS (California Health Interview Survey): healthpolicy.ucla.edu/our-work/california-health-interview-survey-chis
- CleftGeneDB: https://bioinfo.uth.edu/CleftGeneDB/
- CPTAC (Clinical Proteomic Tumor Analysis Consortium): proteomics.cancer.gov/programs/cptac
- Clinicaltrials.gov
- COHRA (Center for Oral Health Research in Appalachia): https://www.dental.pitt.edu/research/center-oral-health-research-appalachia
- Craniorate: www.craniorate.org
- Dataverse: dataverse.org
- dbGaP (database of Genotypes and Phenotypes): ncbi.nlm.nih.gov/gap/
- Dryad: datadryad.org
- EDRs (Electronic Dental Records)
- EHRs (Electronic Health Records)
- EnamelBase: www.facebase.org/resources/enamelbase
- ENCODE (Encyclopedia of DNA Elements): www.encodeproject.org
- Epic Research: www.epicresearch.org
- FaceBase: https://www.facebase.org/
- Framingham Heart Study: www.framinghamheartstudy.org
- GEO (Gene Expression Omnibus): www.ncbi.nlm.nih.gov/geo
- GIS (Geographic Information System) data, including area level measures: www.usgs.gov/the-national-map-data-delivery
- Github: github.com
- GLOBUS for raw data sharing: www.globus.org
- GTEx (Genotype-Tissue Expression) project portal: www.gtexportal.org
- HOMD (Human Oral Microbiome Database): www.homd.org
- IBM Watson: www.ibm.com/watson
- IDC (National Cancer Institute Imaging Data Commons): datacommons.cancer.gov/repository/imaging-data-commons
- IGVF (Impact of Genomic Variation on Function): igvf.org
- Insurance claims data
- KidsFirst: portal.kidsfirstdrc.org
- Medicare/Medicaid databases
- MEPS (Medical Expenditure Panel Survey): meps.ahrq.gov/data_stats/onsite_datacenter.jsp
- MGI (Mouse Genome Informatics): www.informatics.jax.org
- MusMorph: github.com/jaydevine/MusMorph
- National Oral Health Data Portal: www.nationaloralhealthdataportal.net
- NCDB (National Cancer Database): www.facs.org/quality-programs/cancer-programs/national-cancer-database
- NDA (NIMH Data Archive): nda.nih.gov
- NHANES (National Health and Nutrition Examination Survey): www.cdc.gov/rdc/b1datatype/Dt1222.htm

- NHIS (National Health Interview Survey): www.cdc.gov/rdc/b1datatype/Dt1225.htm
- ScHARe (Science Collaborative for Health disparities and Artificial intelligence bias REduction): www.nimhd.nih.gov/resources/schare
- NIH CDE (Common Data Elements) Repository: cde.nlm.nih.gov
- NOHSS (National Oral Health Surveillance System): www.cdc.gov/oralhealthdata/overview/nohss.html
- gnomAD (Genome Aggregation Database): gnomad.broadinstitute.org
- PATH (Population Assessment of Tobacco and Health): https://pathstudyinfo.nih.gov/
- PRESTO (Pregnancy Study Online): www.bu.edu/slone/research/studies/presto
- recount3 - (by ReCount project): rna.recount.bio
- Smart-DOC (Smart Dental Oral and Craniofacial): https://smart-doc.dent.umich.edu/)
- SRA (Sequence Read Archive): ncbi.nlm.nih.gov/sra
- TCGA (The Cancer Genome Atlas): cancergenome.nih.gov
- TCIA (The Cancer Imaging Archive): cancerimagingarchive.net
- Track hubs (with UCSC Genome Browser)
- UCLA (University of California, Los Angeles) Biobank
- UCSC (University of California, Santa Cruz) Genome Browser: genome.ucsc.edu
- UK Biobank: www.ukbiobank.ac.uk
- US Census American Community Survey: census.gov/programs-surveys/acs
- US EPA (Environmental Protection Agency): epa.gov
- VA (Department of Veterans Affairs) Open Data Portal: data.va.gov
- Vivli Center for Global Clinical Research Data: vivli.org

Challenges to Data Sharing

Participants in listening sessions and responses to the RFI listed a range of challenges that they have encountered in sharing data with others, and which in some cases have prevented them from sharing data with others. The list below also includes additional challenges reported by members of the working group.

Technical, Scientific, and Academic Challenges to Data Sharing

- **Size and complexity of data sets:** Large and complex data sets are inherently more challenging to share due to file sizes and data structure.
- **Data storage:** Reliance on internal, institutional storage capacity can be a major challenge, in particular with respect to data sustainability.
- **Uncertainty about utility of data to others:** Data generators may be uncertain which of their data is worth sharing because it may be useful for others.
- **Uncertainty about data formats:** It can be difficult to anticipate what data formats would be most useful for downstream users (e.g., raw or partially analyzed data).
- **Data reuse may require specialized tools or knowledge:** Reuse of data available in repositories may require specialized software, computing equipment and infrastructure, and expertise for analysis and interpretation of data that users may not have.
- **Lack of incentives for data sharing:** Current publication-based incentive mechanisms in academia do not reward data sharing to the extent they should.
- **Strategic considerations conflicting with data sharing:** In many cases, data generators share data only at the time of publication of an associated manuscript to ensure that publication priority is not compromised. This is particularly relevant for protecting and

supporting the careers of graduate students, postdocs, and early career investigators. As another example of strategic considerations conflicting with data sharing, researchers developing AI models may hold on to data for reuse in future AI model development. They may make the code and models open-source and share them quickly but may not share the data used to generate the models, as they are continuously refining, upgrading, and publishing new models based on these same data.

- **Unclear roles and responsibilities:** It can be unclear who within a given research effort is responsible for what aspects of data sharing.
- **Disagreements on data sharing philosophy:** In collaborative research efforts (e.g., consortia), a single participating investigator or institution with the ability to "veto" data sharing may prevent data sharing by the larger group, even when most participating researchers/institutions support data sharing.
- **Lack of suitable repositories:** Some participants reported that they produce types of DOC-relevant data for which no suitable repository currently exists, at least not to their knowledge. They would have to rely on generalist repositories, which may result in lower findability of data and do not generally support direct integration with other relevant data types.
- **Concerns about quality of available repositories:** Some existing repositories do not follow rigorous data quality and metadata standards, which can result in poorly documented data that has limited reuse value.
- **User support by repositories:** Some participants reported lack of user support by some repositories as a hurdle to data upload. Specifically, participants reported lack of responsiveness of a repository despite multiple attempts to get support with depositing data.

Policy and Regulatory Challenges to Data Sharing

- **Human subjects protection:** Data from patients or study participants may involve significant privacy and ethical considerations that may impact the ability to share. Potentially sensitive DOC-specific data types include genetic and genomic data, facial images, and morphometric data. As data science methods advance, there is particular concern that even de-identified data may be potentially identifiable in the future through application of data science methods or integration with other data (e.g., geolocation data). Federal, state, local, or tribal law may also impact the ability to share data. Additional considerations may apply to data from vulnerable populations and participants with rare disorders or conditions.
- **Conditions of "controlled" data use:** "Public" but not "open" data sharing requires data use agreements between data contributors, data repositories, and data users. It often involves limitations or restrictions on how data can be used, determined by the informed consent, and may require formal committee review. When working with specific communities (e.g., American Indian and Alaska Native), additional considerations for indigenous data sovereignty (IDS)[14] and guidelines may apply for involving community members on committees overseeing data reuse access requests.

[14] The CARE Principles for Indigenous Data Governance, available at https://www.gida-global.org/care

- **Lack of advance planning:** Lack of advance planning, such as insufficient consideration of data sharing plans (data storage, controlled data access) for institutional review board (IRB) approval can create significant barriers for data sharing after collection.
- **Granularity of shared data:** To protect participant data, data may be shared as summary results only (e.g., as Genomic Summary Results [GSR]). While reducing the resolution of the data, this approach can broaden access and is sufficient for the research needs of many.
- **Consenting challenges for EHR data:** Patients typically provide consent for use of data for clinical care, but this does not equate to consent to secondary research. It can be challenging to determine what data may be shared appropriately and through what mechanism. Of note, this scenario is different from data resulting from clinical research, where secondary research use is more commonly already considered in the participant consent. It was also noted that this is a particular problem for studies that are collecting data from cohorts over decades, in which early participants were not consented for public sharing of genomic or other data.
- **International regulatory barriers:** Policy hurdles for international data sharing (e.g., EU General Data Protection Regulation, GDPR), may impede or slow the ability to migrate data outside specific regions.
- **Legal uncertainty:** Community members expressed concern that the potential for violating internal policies by sharing data may result in disciplinary action up to regulatory board punishment.
- **Institutional bureaucratic hurdles:** Community members indicated that their institution has slow and tedious processes (e.g., for material transfer agreement [MTA] requests), which creates a significant hurdle to data sharing due to the excessive time commitment that would be required to follow these processes.
- **Proprietary issues:** Proprietary issues may prevent broad data sharing.
- **Company policies:** Company policies, such as management service agreements (MSAs) of companies serving community health organizations, may prohibit data sharing.

Funding and Resource-Related Challenges to Data Sharing:

- **Resources required for sharing data:** A recurrent concern amongst respondents was that data sharing requires significant resources, including staff time and computing resources. In addition, sharing data can require significant expertise and specialized resources that may not be available to a given investigator. While resources for data sharing are increasingly added to NIH grants under new policies, older projects that are already in progress often have no dedicated budget for data sharing.
- **Concerns about long-term sustainability of data and resources created under limited-term funding:** Data sharing activities and responsibilities may continue past the funding period of the data-producing project. Completion of data sharing and maintenance of shared data can be challenging without dedicated resources. This challenge is in part addressed by NIH's new Data Management and Sharing Policies, which requires investigators to provide plans for data preservation and continued access using persistent identifiers.

- **Requirements of specific repositories:** Some repositories have very specific and involved submission requirements. Such requirements may call for extensive documentation and (re-)formatting efforts and thereby exacerbate concerns about the resources required for submission.

Community Perspectives for DOC Data Resources

Listed below are general perspectives and considerations for DOC Data Resources provided by RFI respondents and participants in the listening sessions. These perspectives are provided regardless of consensus agreement by working group members. Please see section "Recommendations" below for consensus recommendations by the working group.

Infrastructure

- **Sufficient capacity for "big data":** Repositories must support very large data storage.
- **Adequate connectivity:** Large capacity must be accompanied by adequate connection to high-speed data resources. Slow connections are prohibitive for downloading large data sets.
- **Consider cloud computing as an option:** For various purposes, researchers are increasingly shifting to cloud-based options to handle very large amounts of data.
- **Persistence:** There are unresolved concerns by many in the community about the persistence of currently existing data resources. Investigators have spent (and continue to spend) significant resources on depositing data into existing repositories, but in some cases their long-term existence and maintenance does not seem guaranteed.
- **Maintenance and updates of data:** As research is completed and data is shared in real time, it is critical that data update mechanisms exist, and their use is incentivized or enforced.

Data Access and Use

- **Open access:** The use of repositories must be affordable, preferably free, for data users.
- **Transparent data use agreements:** Data use agreements should acknowledge the funding source and prohibit users from any attempt to re-identify or to otherwise use the data in an inappropriate way and specify access controls as necessary.

Data Integration

- **Support for data integration:** Integration across data types, and integration with analysis tools is critical, but not available or insufficient in most currently available DOC-relevant data repositories.
- **Integration of experimental and observational data:** Ideal DOC data repositories would facilitate the integration across experimental and observational data sets, including tools for visualization and integrative analysis.
- **Integration across dental and medical data set:** Working towards better integration of dental and medical data is key. This includes, for example, integration of pathology and

radiology with actual imaging and histology data, which can be useful as training data for development of precision medicine approaches.

- **Toward a practice-based research network (PBRN):** There is a need to expand the NIDCR-funded National Dental Practice-Based Research Network (PBRN) to reach the full U.S. population, including low-income and underserved populations.
- **Consider data integration for newly funded projects:** Incorporating data sharing standards into study designs of large projects from their inception requires effort but will substantially increase the value of the resulting data. The 2023 NIH Data Management and Sharing Policy[15] is an important step in this direction.
- **Alignment with NIH-wide efforts:** Any new or expanded existing DOC data repositories need to be carefully aligned and integrated with the large NIH data ecosystem, as well as the NIH Office of Data Science Strategy and the *NIH Strategic Plan for Data Science*[16].

Community Engagement

- **Design repositories with focus on specific user communities:** Some community members commented that repositories often do better when they are very focused and driven towards a specific research community or field. When considering the development of new repositories, it is important to think about what kind of data they can and should house, and ways to reduce barriers to data access.
- **Incorporate community-level information in genomic data sets:** Much genomic and other omics-type data focus on the individual level, with no or limited integration of community- and population-level data. This is a missed opportunity that can potentially be addressed in DOC data resources.
- **Leverage partnerships:** NIDCR should consider fostering collaborations with other federal agencies, organizations, and initiatives working on similar data-related challenges, to leverage their expertise, resources, and lessons learned (see section "Community Leadership for Data Sharing and Reuse" above of examples of relevant organizations and initiative). This can help avoid duplication of efforts, promote the sharing of best practices, and accelerate the development of an effective and robust DOC data infrastructure.

Data Representation, Standards, and Quality

- **Human phenotype data and metadata quality:** The quality of data and metadata in current repositories can vary widely. This was viewed as a special concern for phenotypic data and metadata associated with human genetics data. There are opportunities to address this through development of better standards for human phenotype data and metadata.
- **Uniform data processing:** Especially for genetics data, uniform processing of data that has been produced in non-uniform ways can be challenging. This is currently often pursued by multiple groups reprocessing the same data in parallel. There are opportunities to reduce costs and wasted resources by unified data processing systems.

[15] Data Management and Sharing Policy, available at https://sharing.nih.gov/data-management-and-sharing-policy

[16] NIH Strategic Plan for Data Science, available at https://datascience.nih.gov/strategicplan

- **Data standards:** Beyond human phenotype data, there is a need for better standards for many types of DOC-relevant data.
- **Terminology:** For a meaningful DOC data science strategy, it will be important to develop and appropriately use well-defined terminology describing components of the data ecosystem. Differences in use of terms such as “repository” vs. “computing center” and blurred terminology can undermine meaningful and constructive discussion and planning.

Workforce

- **Training and development:** As new resources are being developed, there is a parallel need to train investigators in their use, especially for sharing and uploading data.
- **Ensure relevant expertise of data managers:** Management of complex data types often requires significant biological and/or clinical expertise to ensure that all required information is adequately included and presented.

General Observations

This section summarizes several general observations about the DOC data ecosystem and related challenges that represent a synthesis of the DSS-WG’s findings from information gathering exercises and from the members’ experience in their own respective fields of research.

Complexity and Heterogeneity

The DOC data ecosystem is very complex and heterogeneous. In the information gathering exercises, more than 60 data systems were identified that are currently being used by DOC researchers. These systems serve various purposes: as information sources, repositories for depositing data, platforms for sharing data with others, or tools for integrative data analysis. Likewise, over 65 different data types were identified that are being used by DOC researchers. Of note, some of these data “types” are not a single type of measurement or data format, but they consist of multiple types of information themselves. Along with this complexity of data types, many data formats and potential data standards are required to capture all of these data types. Despite the DSS-WG’s extensive efforts to collect comprehensive information, this collection of data types and systems is likely incomplete and underestimates the full complexity of the DOC data ecosystem.

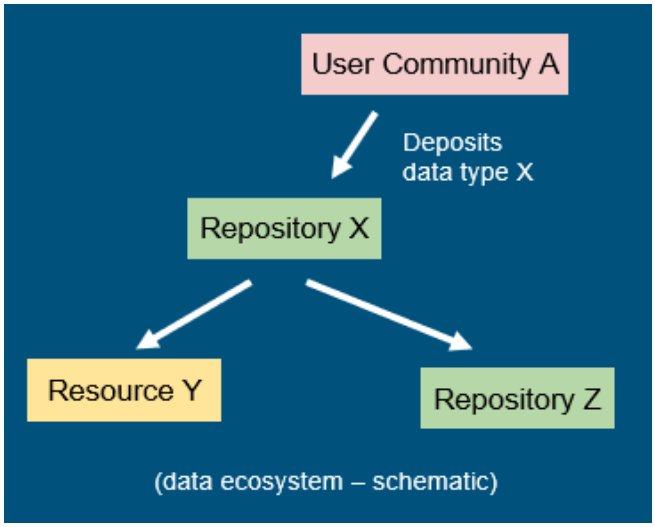

Fig. 1: The DOC data ecosystem is complex and heterogeneous.

Fuzzy Boundaries

The current DOC data ecosystem, as defined by data systems used by DOC researchers, consists of DOC-focused data systems in conjunction with a much larger number of general scientific and medical data systems that contain DOC-relevant data. This lack of delineation is inevitable and often beneficial since it creates synergies across communities and funding agencies in terms of support for these systems. Nonetheless, it can create challenges for the integration and interoperability across systems. In developing an NIDCR data science strategy, strategic alignment between purpose-built DOC-centric systems, other NIH systems, and additional external systems (e.g., those operated by commercial health care providers) is critical.

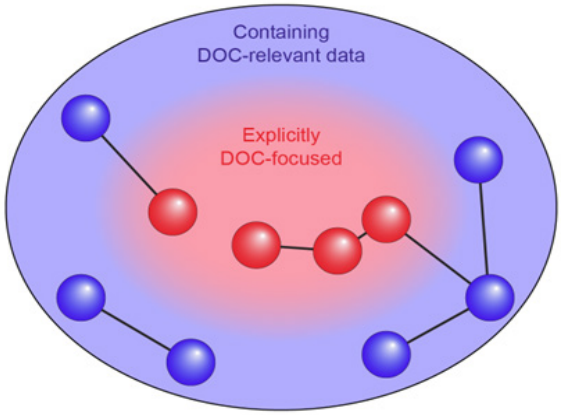

Fig. 2: The DOC data ecosystem has fuzzy boundaries with respect to the larger research data ecosystem.

Lack of Connectivity

The existing DOC data ecosystem is not the result of an intentional design process but has grown organically over decades. Data systems are supported by NIDCR, other NIH institutes, and other public and private funders. Data systems are being created by many different groups and with various degrees of coordination with other DOC and generalist resources. This is reflected in a lack of coherence and interoperability. Many of the data systems used by DOC researchers are not directly connected to each other. It may not be necessary to have all DOC-related data systems connected to each other, especially when they align with research at the opposite ends of the translational spectrum. Although there are also many cases where connectivity would be desirable in principle, the lack of common data and metadata standards, ontologies, and data exchange interfaces remain a challenge. Efforts to enhance connectivity should focus on developing common standards and interfaces to facilitate seamless data exchange.

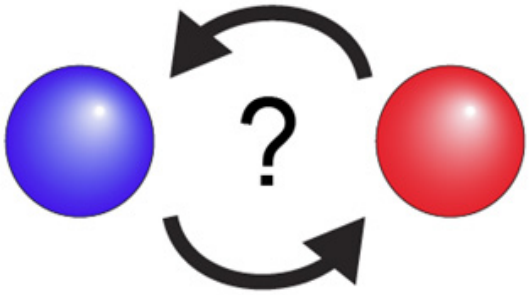

Fig. 3: Within the DOC data ecosystem, resources are not fully interconnected to facilitate data integration.

Lack of Dedicated Resources

A significant gap exists in dedicated resources for supporting diverse data types from NIH-funded projects, including those funded by NIDCR. NIH has made significant strides toward ensuring the sharing of scientific data from NIH-supported research projects by issuing its 2023 *Data Management and Sharing Policy*[17]. However, for many types of data from NIH-funded efforts, including NIDCR-funded efforts, there are currently no obvious existing data systems that provide full support for the respective data type beyond serving as a general data archive. This gap can create conflict with the intent and implementation of the NIH Data Management and Sharing Policy, since researchers cannot share their data in a way that truly facilitates their discovery and reuse by other researchers.

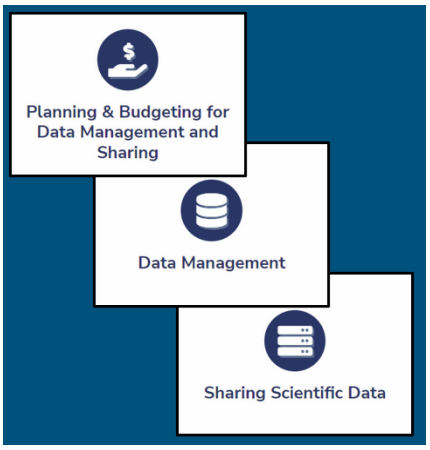

Fig. 4: Resources are needed for building and maintaining data systems and resources.

Funding Instruments can Create Hurdles for FAIR Compliance

The reliance on NIH funding instruments primarily designed for time-limited research projects, such as Research Project Grants (R01) or Research Project Cooperative Agreements (U01), poses challenges for the sustainable operation of data systems and compliance with FAIR principles[18]. Many data systems are currently supported under funding instruments that are primarily designed for research projects of limited duration. The lack of NIH funding opportunities supporting long-term sustainable operation of data systems can result in infrastructure, curation, and governance gaps and hurdles for FAIR data compliance.

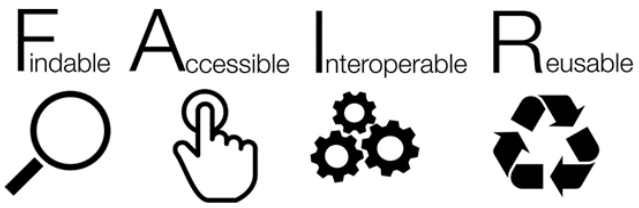

Fig. 5: To be useful, DOC data must be findable, accessible, interoperable, and reusable (image credit: Sangya Pundir, Wikimedia, CC BY-SA 4.0)

[17] Data Management and Sharing Policy, available at https://sharing.nih.gov/data-management-and-sharing-policy

[18] Figure credit: Sangya Pundir, CC BY-SA 4.0 DEED license, available at https://commons.wikimedia.org/wiki/File:FAIR_data_principles.jpg

NIDCR-Specific Opportunities in Data Science

In the following sections, the DSS-WG identifies several opportunities in data science that are specific to NIDCR. Developed by DSS-WG members with relevant expertise, these sections synthesize input from the community and discussions within the working group. The descriptions of these opportunities aim to provide expanded background and a forward-looking vision, rather than specific recommendations. However, they were essential in shaping the Recommendations section of this report, where the DSS-WG outlines specific steps to realize these opportunities.

NIDCR Data Science and Oral Health Disparities

Oral health disparities arise from systemic interactions between biological, psychological, social, environmental, economic, policy, and other factors. While there is increasing awareness of these complexities, the study of oral health disparities has historically focused on disaggregated analyses of individual factors in isolation. Such disaggregation of the complex systems responsible for the emergence of disparities has left us with an incomplete and inadequate understanding of the causal mechanisms, often leading to the implementation of ineffective population-based interventions. The predominance of disaggregated investigations is driven primarily by three separate but related problems:

1) The absence of a well-integrated data ecosystem that can be leveraged to address the complex causes of oral health disparities,

2) The lack of comprehensive, high-quality, and unbiased data for integration due to underrepresentation of necessary populations and lack of standardized key data elements, and

3) The lack of a workforce capable of developing and using tools and methodologies that can effectively leverage ‘big data’ and integrate different data types.

The emergence of data science, driven by the development of novel analytical methods, advances in computational capacity, and the creation of new data systems and repositories, provides tools with significant potential for reducing oral health inequalities in the coming decades. To unleash the full potential of data science for identifying and mitigating oral health disparities, diversity and oral health inequalities should be considered the central topics when envisioning the future DOC data ecosystem. Challenges and opportunities identified by the DSS-WG are outlined below.

Integrate social determinants of health into DOC data systems, emphasizing disadvantaged and underserved populations. Data diversity and integration provide challenges in all areas of DOC research, as described in detail in other sections of this report. The consequences of these challenges are particularly pertinent in the area of oral health disparities, underscoring the pressing need for targeted data generation activities. When such data is integrated into the broader data ecosystem, a more complete rendering of oral health disparities and distributions is possible. Establishing standardized, valid, and reliable domains and measurement tools to capture social determinants of health across DOC data repositories and systems is critical to realizing the true potential of data science to reduce oral health disparities and inequalities. These measures can be combined with other complex, multi-level,

multi-sector repositories to assess holistically the root causes of oral health disparities. One important opportunity arising from these efforts will be the identification, characterization, and quantification of potential data asymmetries (i.e., the disproportionate paucity of data for minorities, underserved populations, and other disadvantaged groups compared to the rest of society). Understanding such asymmetries in detail will be invaluable for creating more equitable data sets in the future. The use of data tools and methods capable of leveraging data diversity will be helpful in identifying such critical gaps in data elements and depth of data needed to better understand how and why oral health disparities and inequities occur.

Combine data from DOC-specific and general data ecosystems to better understand the fundamental reasons behind inequalities in oral health. A diverse DOC data ecosystem, in conjunction with other data ecosystems, can provide data to discover and assess complex interactions across social determinants of health, biological and psychological factors, and environmental variables. When combined with public health monitoring systems and observational data, a well-integrated data ecosystem can be used to evaluate the impact of national and subnational oral health policies, policy experiments, and national emergencies and their influence on the US population's oral health. Contextualizing data by geography or population characteristics can pinpoint specific oral health needs and inequities. These findings can serve as a starting point to design interventions at a clinical and population level based on the underlying root causes of social needs and health inequalities. Additionally, such efforts also have the potential to reveal new and unexpected predictors of poor oral health. By integrating various data sources and applying advanced analytical methodologies, researchers may achieve comprehensive insights into oral health patterns. This includes a deeper understanding of the impact of social determinants, lifestyle factors, and behavioral aspects on oral health. In addition, identifying factors closely linked to these disparities could help recognize areas prone to oral health inequalities, and guide the development and implementation of new preventive policies. Combining data from different ecosystems can pose significant technical challenges regarding data integration and harmonization of metadata. However, if these challenges can be overcome, a well-established and interconnected DOC data ecosystem has the potential to fill data gaps for racial and ethnic minorities and other underserved populations that have been consistently excluded from clinical trials, surveys, and other study designs.

Enhance applications of data methodologies to better leverage the DOC data ecosystem for novel insights into oral health disparities. Complementary to an interconnected DOC data ecosystem, new methodological approaches are needed to leverage the complexity of data types to advance the study of oral health disparities and to assist in the design and study of multi-level interventions aimed at addressing such disparities. Complex systems science approaches incorporate computational and machine learning approaches to develop data-informed dynamic computer simulations capable of generating the emergent properties of a system. Such simulation methods can facilitate the assessment of novel and complex interventions to tackle oral health inequalities, including scalability, effectiveness, and long-term sustainability. For example, cost-effectiveness analyses using DOC data can be used to inform public health and population-level interventions, optimize the use of resources, increase coverage, and further close oral health inequalities. These methodologies may also serve as valuable tools in highlighting the limitations

of the current DOC data ecosystem as it relates to the investigation of oral health disparities and inequities.

Support the development of data science knowledge and application in DOC population health researchers. The current state of the science in oral health disparities also serves to highlight the importance of a DOC research workforce that is trained in the data sciences beyond population methods. Currently, there is no dedicated DOC funding opportunity that provides research training in population oral health or which has a specific focus on data science, both of which are needed to adequately investigate oral health disparities. While cross-disciplinary collaboration can serve to address some aspects of this issue, effective collaboration between DOC researchers and data scientists can prove challenging without at least some training on the part of the DOC researchers. Developing a pipeline of DOC population health researchers with data science experience or expertise to provide a foundation to address oral health disparities would have remaining challenges. The lack of an integrated and available data ecosystem can limit research training opportunities, but also cannot be addressed without a well-trained and capable workforce committed to improving DOC health and overall health.

Close monitoring of artificial intelligence (AI) applications for DOC data to minimize unfairness and bias in clinical and policy decision-making. AI and machine learning (ML) hold major potential for applications in DOC data science (see section “Artificial Intelligence and Machine Learning Readiness of Data” below). However, algorithms constructed using DOC-specific and related data must be based on datasets that are free of racial-ethnic and socioeconomic biases. Such biases of AI systems have been well documented in other sectors and, if not addressed, have potential to occur in DOC and other health data due to structural inequalities in health systems and research, which may distort the data that is used to build and test (e.g., prediction models). The consequences of using biased models are the perpetuation and potential magnification of past inequalities into future decision-making. Thus, these models should be tested before deployment and monitored after implementation to minimize the chance of harm.

Artificial Intelligence and Machine Learning Readiness of Data

Artificial intelligence (AI) and machine learning (ML) constitute an ensemble of data-driven technologies with major potential for enhancing biomedical research and healthcare practices. Notable areas of applications relevant to DOC research include the integration across multimodal data types for basic discovery science; the analysis of very large genetics and genomics data sets; advanced natural language processing (NLP) approaches for the analysis and interpretation of scientific literature, patient records, or any other text-based information sources; analysis and annotation of clinical 2D- and 3D-imaging data to assist in detection and diagnosis; advanced analyses of public health data and integration of health-related information with other complex data sets (e.g., demographic, socioeconomic, geographical); predictive modeling for individual patient outcomes; development of personalized treatment plans; optimization of dental devices, implants and regenerative medicine scaffold designs as well as materials selection based on patient-specific criteria; analysis of patient behavioral data to understand compliance and disease progression; and enhanced tele-dentistry, especially for diagnostics and patient engagement.

However, the effectiveness of AI/ML is intricately linked to the readiness of the biomedical data being analyzed by these systems. First and foremost, AI/ML-ready health data must adhere to the FAIR guiding principles for scientific data management and stewardship, emphasizing its findability, accessibility, interoperability, and reusability. The preparation of biomedical data to ensure optimal AI performance entails critical steps and considerations described below.

Establishment of data governance and equitable data curation. A well-defined and robust data governance and curation framework is crucial for the efficient management, protection, and utilization of data. This requires establishing clear policies and procedures for data management, access controls, application and data ontologies, standards to improve interoperability, and data lifecycle management. It is also necessary that all organizations, groups, and individuals involved in DOC research understand and adhere to these governance principles. The framework also needs to address intellectual property protection of information, HIPAA compliance, protection of personal information, and auditing processes.

Assurance of data quality. Successful application of AI/ML models relies heavily on accurate and high-quality data. This requires the establishment of data quality metrics through data cleaning and normalization, patient-centric labeling, and annotation techniques. It also requires the implementation of processes to maintain and improve data quality over time to ensure the regular monitoring, validation, and cleaning of data to keep it reliable and current. Data cleaning involves identifying and rectifying errors, removing duplicates, imputing missing data and incomplete modalities, and removing spurious artifacts from heterogeneous data sources. Normalizing data ensures consistency by standardizing units, scales, and formats, facilitating better model performance.

Data integration and fusion. AI/ML systems often require data from multiple diverse sources and modalities, such as structured and unstructured text, radiographs, pictures, and genetic or genomic information, which may be sourced from different healthcare providers across disciplines. Implementation of effective data integration and multi-modal fusion strategies needs to seamlessly combine data from various platforms and formats, addressing harmonization of distributed or federated data for distributed or federated learning, and association across different modalities and longitudinal studies.

Scalability for volume and velocity. Robust and accurate AI/ML models often require large volumes of data for training and updating, must handle data in real-time, and accommodate multi-scale applications. Data infrastructure needs to be designed to scale up for both volume and velocity, ensuring that it can accommodate the growing demands of AI applications. In this context, AI/ML approaches have significant potential for the discovery, identification, and mitigation of errors and imbalances in the data and biases in data labels or metadata.

Data security and compliance. AI/ML readiness demands a robust data security and compliance framework. AI/ML systems must safeguard sensitive information, comply with relevant regulations (e.g., GDPR, HIPAA), and ensure that data is handled responsibly throughout its lifecycle. They also need to establish protocols for data access, encryption, and anonymization to protect both user privacy and organizational integrity. The critical need for data security and compliance with relevant regulations requires the implementation of robust protocols for data access, encryption, and anonymization to safeguard sensitive information.

Converting Data into Applications

The dental and oral health communities can markedly advance translational medicine and therapeutic development by leveraging data science innovations. Such data-driven innovative approaches have already resulted in FDA-approved devices and therapeutic strategies across various medical domains. Below are key areas where data integration is reshaping patient care:

Strategies for personalized therapeutics. The transformation of data into therapeutic solutions might take the form of data-driven molecule or drug development, digital biomarker development, and stratification of clinical risks. This process utilizes data derived from patient records, plain radiographs, 3D imaging, direct imaging, pathology, microbiomes, and laboratory test results.

Clinical record analytics. Data science enables the extraction of clinically meaningful insights from patient records through techniques such as large language models (LLMs) and natural language processing (NLP), data mining, and machine learning algorithms applied to electronic health record data. These methods can be used in healthcare to extract and summarize information from clinical notes, discrete laboratory data, and other free-text documents. In medicine, electronic health record data has been mobilized to predict the risk of an individual's need for transfer into an intensive care unit. In dental and oral health applications, predictive analytics could identify patients at risk for various diseases based on patterns in their patient records. Analytics on aggregated dental records can reveal useful trends in oral diseases and treatment responses across patients. Furthermore, patient dental records that suffer from a lack of completeness can be supplemented through data analysis of these trends.

Diagnostic imaging enhancement. Plain film x-rays including panoramic x-ray (panorex) studies can be analyzed with data science techniques to assist dental healthcare professionals in making accurate diagnoses. Throughout healthcare, radiographs suffer from imperfect interpretation or reproducible technical errors from data capture. Using computer vision to aid in the interpretation of panorex studies, as well as evaluating their diagnostic necessity in the first place, can reduce the number of radiographs needed per patient, thereby reducing the amount of radiation a patient is exposed to. Techniques such as noise reduction can also improve direct image quality for more accurate interpretation, and techniques such as uncertainty quantification can provide clinicians with understanding of the reliability of a given study.

3-D imaging for precision medicine. Data science approaches can precisely segment computed tomography (CT) and magnetic resonance imaging (MRI) data, which is crucial for a variety of applications in DOC care. These techniques have been instrumental in oral cancer management, allowing for the precise segmentation of organs and lesions and the quantification of their volumes, densities, and changes over time, facilitating both planning and tracking of treatment. These techniques are also used to evaluate bone for dental implant placement and difficult tooth extractions to avoid possible complications during and after surgical procedures. Additionally, these data science approaches provide invaluable insights into disease progression, evaluation of temporomandibular joint disorders, aid orthodontic assessments, and facilitate airway analysis in cases of obstructive sleep apnea. They also play a critical role in mapping structures for reconstructive surgery, diagnosing endodontic issues, identifying craniofacial pathologies, and guiding facial trauma recovery. The integration of computer vision enhances the diagnostic utility of 3D images, enabling the alignment of CT and MRI images with each other or

with radiographs, thereby consolidating diagnostic information. This alignment is particularly beneficial for accurately tracking disease progression over time. Automated segmentation of anatomical structures, such as individual teeth, can further refine diagnostic and treatment planning processes. Overall, advances of data science in 3D imaging diagnostics have the potential to revolutionize the field of DOC health. It enables virtual surgical planning, integration with computer-aided design/computer-aided manufacturing (CAD/CAM) systems for prosthetic design, monitoring of craniofacial growth and development, assessment of sinus anatomy, visualization of vascular and nerve layouts, and understanding of soft tissue structures for grafting or surgical interventions.

Rapid diagnostic imaging. Direct imaging scans can be analyzed automatically and instantly with data science to assist in expedited diagnoses and treatment planning. Direct imaging and digital dental radiographs are used in dentistry to detect cavities, evaluate cancer risk in suspicious oral mucosal lesions, and to check the status of developing teeth and monitor tooth and bone health. Using computer vision and deep learning, scientists can automate the segmentation of anatomical structures in direct imaging to assist in diagnosis and surgery planning. AI techniques can also reduce noise, increase resolution, adjust contrast, magnify, and clarify direct images.

AI-supported pathology diagnosis. Data science enables more accurate and expedient pathology diagnoses through computational analysis of tissue imagery and leveraging machine learning algorithms trained on comprehensive patient biopsy data. In dental and oral health, pathology is used to examine biopsied tissue to diagnose oral diseases. Deep learning, a type of ML based on artificial neural networks, has been used to classify tissue as cancerous, as well as identify molecular disease subtypes such as specific mutation classes. Additionally, generative deep learning techniques can be used as an educational tool to further oral health pathologists' understanding of disease pathobiology. Finally, as image data becomes increasingly multiplexed with emerging technologies, data science can incorporate this challenging, intercalated data into a single streamlined clinical tool.

Microbiome informatics. Data science enables a deeper understanding of the role of oral and gut microbiomes in health and disease through genomic sequencing, bioinformatics analysis, and examination of microbial data sets using machine learning. Microbial DNA can be sequenced to understand the diversity, relationships, and abundances of microbes in a microbiome community. Resources such as the expanded Human Oral Microbiome Database (eHOMD) can be applied to investigate the microbial causes of periodontal disease with artificial intelligence-based complexity reduction tools.

Generative technologies. Generative technologies can be trained from massive databases of molecular, basic science, and clinical disease data to create clinically useful *de novo* molecules. Indeed, AI research teams have rapidly proposed, synthesized, and tested new molecules for human diseases[19], with early concepts of generatively designed molecules now in phase 2 clinical

[19] Zhanoronkov et al. (2019), *Deep learning enables rapid identification of potent DDR1 kinase inhibitors*, Nature Biotechnology 37:1038-1040, doi: 10.1038/s41587-019-0224-x

trials[20]. Dental and oral health researchers might use generative techniques to curate favorable microbial colonies, design new adhesives, or reverse premalignant progression.

Leveraging Data Diversity

A distinguishing feature of DOC research is the diversity of data types and experiments. During the listening sessions conducted in preparation of this report, data diversity was mentioned repeatedly. **Table 6** (above) shows a summary of the data types enumerated in listening sessions that were conducted to get community feedback on DOC data issues.

As highlighted in the listening sessions, the diversity in DOC data types is substantial. As an example, the FaceBase data repository, which has only recently expanded its scope to encompass a broader range of DOC data currently integrates data with over 47 different experiment types, which include genomics, imaging, population health, and secondary analysis. Such diversity is undoubtedly a strength because it offers multifaceted opportunities in basic DOC research and has the potential to generate unique insights into oral health. However, it also presents challenges. It necessitates strategic integration to prevent data from becoming siloed and to maximize the potential of cross-data analytics. The diversity of DOC data requires strategies for data management and analysis. The subsequent sections will explore the specific challenges posed by this diversity, as well as the numerous opportunities it presents for advancing the field.

Challenges Arising from Data Diversity

Integration and value extraction. Integrating diverse DOC data types creates significant challenges. As an illustrative example, a recent study of regulatory mechanisms of palate development[21] used a combination of data types from mouse embryonic palate tissue available through FaceBase including multi-omic single-cell data, imaging data, and spatial transcriptomics data. Systematic integration across such data types requires shared metadata standards for effective linkage. This example highlights the need for a strategic approach to integrating various data types to avoid data siloing and enhance the potential for cross-data analytics.

Repository diversity and data fragmentation. The DOC data ecosystem comprises both specialist and generalist repositories. Specialist repositories like ENCODE, GEO, and dbGaP are suitable for certain subsets of data but may not cover the entire spectrum of data types in DOC research. This can lead to data fragmentation across multiple repositories, making it challenging to establish connections between different types of data.

Data utility and operational efficiency. A great diversity in experiment and data types also poses significant challenges for data repositories. Obtaining value from aggregate data collections demands that the data be integrated at some non-trivial level in order to make connections from one data set to another. Higher data utility and more cost-effective operations can be achieved

[20] *Study Evaluating INS018_055 Administered Orally to Subjects With Idiopathic Pulmonary Fibrosis (IPF)*, clinical trial information available at https://clinicaltrials.gov/study/NCT05938920

[21] Piña JO et al., *Multimodal spatiotemporal transcriptomic resolution of embryonic palate osteogenesis* (2023). Nature Communications 14:5687, doi: 10.1038/s41467-023-41349-9.

as the issues of diverse data types are considered as an underlying requirement of a data science strategy.

Challenges with generalist repositories. The NIH has defined an overall data repository strategy[22] in which narrow specialist repositories have a limited range of data types, with the advantage that detailed metadata models may be obtained. On the other side of the spectrum, so-called generalist repositories are currently being recommended for unique heterogeneous collections of data. However, the open and unstructured nature of generalist repositories such as Figshare and Zenodo also result in fragmentation and suffer from lack of uniformity and minimal standardization across data contributions.

Opportunities Arising from Data Diversity

Comprehensive research insights. Integrating diverse data types enables a holistic understanding of oral health conditions, examining interactions among various factors to understand oral health disparities.

Innovative analytical methods. Diversity in DOC data allows for the application and development of novel data science methodologies, such as complex systems science, machine learning, and computational simulations.

Enhanced predictive modeling. Leveraging diverse data types improves predictive models for more accurate prediction of treatment outcomes and disease progression, which has the potential to benefit personalized treatment plans and preventive oral health care.

Data integration for policy and public health. Combining DOC data with broader health data sets offers insights into oral health policies and interventions, helping to customize policies for specific oral health needs and inequities.

Training and workforce development. Managing diverse data types highlights the need for a skilled workforce in DOC-specific data science methodologies, extending to educational programs focusing on data integration and analysis.

Ethical AI in DOC research. Analyzing diverse DOC data sets with AI and ML offers the chance to develop and apply ethical AI principles, ensuring models are free from biases and rigorously tested for accuracy and fairness.

Improved data management and accessibility. A unified data ecosystem capable of handling diverse DOC data types enhances data accessibility and interoperability, aiding research and clinical decision-making.

[22] Martone, M., & Stall, S. (2020). NIH Workshop on the Role of Generalist Repositories to Enhance Data Discoverability and Reuse: Workshop summary. National Institutes of Health, Office of Data Science Strategy. Available at https://datascience.nih.gov/data-ecosystem/NIH-data-repository-workshop-summary

Recommendations

In this report, the DSS-WG compiled detailed information about DOC data science and the DOC data ecosystem, combining the expertise of its members across the translational spectrum with input received from the broader DOC research community. From this wealth of information, the DSS-WG synthesized a set of specific recommendations to guide the future of NIDCR's data science initiatives.

Given the growing importance of data science in DOC research, and the major opportunities in DOC research for applications of data science, the DSS-WG recommends for the **NIDCR to develop a dedicated data science strategy**. This strategy should apply the goals and objectives of the *NIH Strategic Plan for Data Science*[23] to the vision and mission needs of NIDCR, as described in the *NIDCR Strategic Plan.*[24] To facilitate alignment, the Working Group's recommendations are organized using the framework of goals provided in the *NIH Strategic Plan for Data Science*, which involve 1. the creation of a framework to support and enhance data infrastructure; 2. modernization of the data ecosystem; 3. development of advanced tools for data management and analysis; 4. support for workforce development; and 5. establishing policies for data stewardship and sustainability. An updated *NIH Strategic Plan for Data Science* is currently under development and selected preliminary goals from the current draft plan[25] have been incorporated into the recommendations.

Given the wide range of challenges that will need to be addressed and the substantial investments that will be required for doing so, NIDCR should consider leveraging existing resources, initiatives, and efforts when possible, redirecting activities as appropriate to align more closely with the DSS-WG's recommendations.

1. Establish a Robust Data Infrastructure Tailored for DOC Research and Interfacing with NIH Data Systems

1.1 Optimize Data Storage and Security: Develop a data infrastructure optimized for the large volume and diverse data types specific to DOC research, including varied formats such as 3D imaging, microbiome data, genomic data, and clinical data (EHR). This may include cloud-based infrastructure. Ensure such infrastructure incorporates advanced security measures to protect sensitive patient data, adhering to Health Insurance Portability and Accountability Act (HIPAA) and other relevant standards.

1.2 Connect DOC-specific and General NIH Data Systems: Create an integrated DOC research data network that connects with the NIH Data Commons and aligns with NIH's vision of a federated biomedical research data infrastructure. Emphasize interoperability standards like FHIR (Fast Healthcare Interoperability Resources) to enable seamless data

[23] NIH Strategic Plan for Data Science (2018-2023), available at https://datascience.nih.gov/strategicplan

[24] NIH Publication No. 22-DR-8175, available at https://www.nidcr.nih.gov/sites/default/files/2022-01/NIDCR-Strategic-Plan-2021-2026.pdf

[25] December 2023 draft of updated NIH Strategic Plan for Data Science (2023-2028), available at https://datascience.nih.gov/sites/default/files/NIH-STRATEGIC-PLAN-FOR-DATA-SCIENCE-2023-2028-final-draft.pdf

sharing and leverage the potential of EHRs and other real-world data, fostering a culture of data sharing consistent with FAIR principles. This network should facilitate seamless sharing of data and DOC-relevant software tools across various research domains and support collaboration between dental researchers and other health disciplines to foster integrative studies on systemic health and oral diseases. To enable and facilitate studies of oral health disparities, DOC data systems should either incorporate or seamlessly interface with systems containing complex data (e.g., demographic, socioeconomic, geographical) that can be used to contextualize oral and general health data, reveal skews in availability of data from underrepresented populations, and provide a foundation for developing effective individual- and population-level interventions.

1.3 Ensure Ethical AI/ML Readiness of Data Infrastructure for DOC Research[26]: Prioritize initiatives that support, facilitate, and, where necessary, require that DOC data be AI/ML-ready. These efforts should include adherence to the FAIR guiding principles and the development of data governance frameworks, establishment of quality assurance protocols, creation of integration strategies, design of scalable solutions, and implementation of security measures. These actions, outlined in the "Artificial Intelligence and Machine Learning Readiness of Data" section of this document, are crucial. Collaboration to bridge AI/ML technical gaps, standardization of data collection, and the seamless fusion of DOC data from diverse sources are essential steps toward this readiness. This may include leveraging other existing AI/ML efforts at NIH[27] for the development of DOC-specific AI/ML infrastructure. Any newly developed AI/ML approaches for DOC research should be carefully tested and continuously monitored for possible biases that may arise from racial-ethnic, socioeconomic, or other underrepresentation biases in the underlying training data.

2. Modernize Data Ecosystems Specific to DOC Research

2.1 Establish Dedicated Data Repositories for DOC Research That Reflect NIH's Objective for Cost-effective, Sustainable, Secure, and Accessible Data Repositories: Establish a repository or a set of repositories for DOC research that integrates data from various research domains such as clinical studies, social studies, epidemiological surveys, and basic science. This system should be designed to support data harmonization and foster standards that enable interoperability with other biomedical databases, facilitating a comprehensive approach to oral health research. NIH grant awardees have the responsibility to contribute their data for public sharing and the ecosystem should provide tools and data workflows that support DOC researchers in meeting the obligation to share data associated with NIH funding. The ecosystem should be sufficiently resourced and

[26] Note: This recommendation (1.3) extends beyond the framework of the 2018-2023 NIH Data Science Strategy. The inclusion of AI data readiness and ethical AI use in our recommendations acknowledges the rapidly evolving field of data science since 2018. With the increasing integration of AI and ML in research, it is imperative to anticipate and align with emerging trends and principles that will likely be emphasized in the forthcoming NIH Data Science Strategy (expected in 2024). This addition underscores our commitment to addressing the challenges and ethical considerations that come with these advancements, ensuring that our recommendations remain current and responsible.

[27] Artificial Intelligence at NIH: https://datascience.nih.gov/artificial-intelligence

organized to support the download of large volumes of data in feasible time frames or support the effective analysis of data in cloud-based systems.

2.2 Support the Storage and Sharing of Individual Data Sets and Software Tools: Implement a platform that allows individual DOC researchers to upload, share, and link their data sets to publications. This platform should promote the use of standardized DOC data elements and metadata to improve data discoverability and reusability. NIDCR should encourage and, where appropriate, require the use of open, non-proprietary data formats to facilitate data sharing.

2.3 Leverage Ongoing Initiatives to Better Integrate Clinical and Observational Data into DOC Data Science: Utilize data from NIH-wide initiatives such as the *All of Us* research program, and as many as possible of the 60 different data systems, databases, data resources, and websites to which DOC researchers currently contribute data or which they use in their work as a source of data or information (see section on *Data Resources and Repositories Used in DOC Research* in this document) to enrich DOC data sets. Encourage the incorporation of oral health data sets into comprehensive data science analyses of longitudinal and life-course studies to elucidate the correlations and impacts of oral health on overall well-being, risk profile, and disease progression.

3. Foster the Development of Data Management, Analytics, and Visualization Tools for DOC Research

3.1 Support Useful, Generalizable, and Accessible Tools and Workflows: Develop and support tools that align with NIH's strategic objectives to generate FAIR data that adds value to research investments and are specifically designed for the analysis of complex DOC data types, such as periodontal charting data, orthodontic cephalometric data, 3D image analysis, morphometry data from model organisms, or oral microbiome sequences. Encourage and support community efforts to develop and systematically apply data standards, terminology standards, and ontologies for DOC-specific data types. These data standards should be compatible with existing standards for widely used data types where applicable. Support the development of innovative systems-level approaches for identifying and addressing the causes of oral health disparities from complex data sets combining DOC data with complementary demographic, socioeconomic, geographical, and other data types.

3.2 Broaden the Utility, Usability, and Accessibility of Specialized Tools: Support the adaptation and refinement of advanced analysis tools from a variety of medical and basic research fields for broader application in DOC research, ensuring that tools are compatible with community standards for data sharing and reusability. This includes advanced imaging analysis tools from related medical imaging fields, visual analytics tools to support clinical epidemiology and population health for cohort identification, integration of bioinformatics technologies for genetic analysis, AI-based diagnostic systems, and advanced materials science for prosthetics and implants. Such adaptations and the development of new tools derived from them, incorporating AI/ML technologies to elucidate biological processes, would enable new clinical treatments and diagnostic technologies into a range of DOC conditions.

3.3 Improve Discovery and Cataloging Resources: Support the development of robust search and discovery tools to enhance the findability and usability of DOC data, thereby improving the value and impact of the data resources.

4. Enhance Workforce Development in Data Science within the DOC Research Community

4.1 Further Strengthen the NIDCR-internal Data Science Workforce: Develop or provide access to data training programs for NIDCR staff and recruit data scientists and others with relevant expertise for internal research efforts and program management. Develop a comprehensive support system for researchers by establishing a 'Data Science Help Desk' within NIDCR. This service would provide expert advice and assistance in the early stages of data collection, particularly in the standardization and annotation of data with Common Data Elements (CDEs). A knowledgeable team in the 'Data Science Help Desk' would be available to guide researchers through best practices in data management and facilitate the adoption of standards, enhancing data quality and interoperability.

4.2 Develop a Strong and Diverse Future DOC Data Science Research Workforce: Prioritize inclusive training programs within the DOC field, with the strategic goal of enhancing human-derived data for research. Commit to recruiting and supporting trainees from underrepresented groups, including racial and ethnic minorities, individuals from socioeconomically disadvantaged backgrounds, and women in data science. Ensure that the training not only covers practical aspects of data management with a focus on standardization and the use of common data elements but also promotes an environment of cultural competence and diversity awareness. Create opportunities for targeted training programs for data science within the dental and craniofacial fields, focusing on specific challenges such as the management of longitudinal clinical data, the analysis of high-dimensional biological data, and the integration of oral health data with population--level non-medical data types (e.g., demographic, socioeconomic, geographical) for studies of oral health disparities. Expand the research workforce in the DOC field by integrating data science into educational pathways and career development programs. Provide funding opportunities to support the training of the next generation of data scientists/clinicians in the DOC arena.

4.3 Engage a Broader Community: Establish a data science unit within NIDCR charged with the role of engaging a diverse community of academic and industry researchers, clinicians, and scientists in data-driven DOC research through collaborative projects and educational initiatives. This unit would provide clarity and support to external users and contributors, ensuring that the NIDCR data ecosystem is accessible, user-friendly, and efficiently utilized for advancing oral health research. This would include information and resources on the NIDCR web site and outreach efforts at relevant scientific meetings and conferences to inform the community about available resources, tools, and data-related opportunities. Encourage industry to leverage DOC data for the development and evaluation of FDA-cleared products and allow clinicians to use it via commercial channels to make DOC research more sustainable.

5. Promote Stewardship and Sustainable Data Policies in DOC Research

5.1 Develop Policies for a FAIR DOC Data Ecosystem: Develop and implement data governance policies that ensure the integrity, confidentiality, and FAIRness of data in DOC research, in line with NIH's strategic goal to improve data management and sharing. Establish a means to support existing and newly created DOC data infrastructure that is suitable to ensure the long-term sustainability of these data resources.

5.2 Enhance Stewardship of DOC Data: Establish stewardship guidelines that emphasize the importance of data quality, utility, and efficiency and define the lifecycle of DOC research data, including retention, archiving, and when necessary, the purging of obsolete data, with particular attention to rare DOC conditions and long-term epidemiological studies. This can be achieved by standing up a data policy committee coordinated by NIDCR, supporting NIH's strategic plan to advance robust data governance frameworks and cross-disciplinary collaborations.

Appendices

Appendix 1: Contributors

Roster

Axel Visel, PhD (Chair) – Senior Scientist, Lawrence Berkeley National Laboratory; Deputy of Science, Joint Genome Institute; Adjunct Professor, University of California, Merced
Amit Acharya, BDS, MS, PhD, FAMIA (Co-Chair) – President, Advocate Aurora Research Institute; Chief Research Officer & System Vice President, Advocate Aurora Health
Lynn M. King, PhD (Co-Chair) – Director, Division of Extramural Activities; Executive Secretary, National Advisory Dental and Craniofacial Research Council, National Institute of Dental and Craniofacial Research

Alexander T. Pearson, MD, PhD – Director of Data Sciences; Director of Head/Neck Cancer Program; University of Chicago, Section of Hematology/Oncology

Alonso Carrasco-Labra, DDS, MSc, PhD – Associate Professor, Center for Integrative Global Oral Health; Director, Cochrane Oral Health Collaborating Center; School of Dental Medicine, University of Pennsylvania

Amy Slep, PhD –Professor and Co-Director, Family Translational Research Group, New York University

Brenda Heaton, PhD, MPH – Associate Dean for Research, Associate Professor of Epidemiology and Health Services Research, University of Utah School of Dentistry; Adjunct Associate Professor, Population Health Sciences, University of Utah School of Medicine

Carl Kesselman, PhD – William H. Keck Professor of Engineering, Epstein Department of Industrial and Systems Engineering; Director, Informatics Systems Research Division, Information Sciences Institute; Viterbi School of Engineering; Professor, Department of Population and Public Health Sciences, Keck School of Medicine; Professor, Biomedical Sciences, Ostrow School of Dentistry; University of Southern California

Fleming Y. Lure, PhD – Chief Product Officer, MS Technologies Corp

Lucia Cevidanes, DDS, MS, PhD – Thomas and Doris Graber Endowed Professor of Dentistry; Department of Orthodontics and Pediatric Dentistry; University of Michigan, School of Dentistry

Melissa Clarkson, PhD, MDes, MA – Assistant Professor, Division of Biomedical Informatics, University of Kentucky

Stefano Monti – Associate Professor of Medicine, Biostatistics, and Bioinformatics; Section of Computational Biomedicine; Boston University Chobanian & Avedisian School of Medicine

Vance Bauer, MA – Vice President of Research; OCHIN, Inc.

Ex Officio Members and Significant Contributors

Alicia Chou, MS – Health Specialist, Division of Extramural Research, Translational Genomics Research Branch, National Institute of Dental and Craniofacial Research

Hiroko Iida, DDS, MPH – Director, Oral Health Disparities and Inequities Research Program, National Institute of Dental and Craniofacial Research

Noffisat Oki, PhD – Director, Data Science, Computational Biology and Bioinformatics Program, National Institute of Dental and Craniofacial Research

John Prue, MS – Technology Officer, Office of the Director, National Institute of Dental and Craniofacial Research

Lu Wang, PhD – Senior Advisor of Data Science, Office of the Director; Chief, Translational Genomics Research Branch; Director, Translational Genetics and Genomics Program, Division of Extramural Research; National Institute of Dental and Craniofacial Research

Other Contributors

Cristina Williams, BA – Consortium Coordinator and Communications Specialist, FaceBase; Information Sciences Institute, University of Southern California

Appendix 2: List of Abbreviations

AAOF	American Association of Orthodontists Foundation
ADA	American Dental Association
AI	artificial intelligence
API	application programming interface
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
BAM	Binary Alignment Map
BRFSS	Behavioral Risk Factor Surveillance System
BRIDG	Biomedical Research Integrated Domain Group
Bridge2AI	Bridge to Artificial Intelligence
BWHS	Black Women's Health Study
CAD	computer-aided design
caDSR	Cancer Data Standards Registry and Repository
CAM	computer-aided manufacturing
CBCT	cone beam computed tomography
CDEs	Common Data Elements
ChIP-seq	chromatin immunoprecipitation sequencing
CHIS	California Health Interview Survey
COHRA	Center for Oral Health Research in Appalachia
CPT	Current Procedural Terminology
CPTAC	Clinical Proteomic Tumor Analysis Consortium
CT	computed tomography
CTCAE	Common Terminology Criteria for Adverse Events
dbGaP	database of Genotypes and Phenotypes
DCN	Data Curation Network
DOC	dental, oral, and craniofacial
DOI	digital object identifier
DSS-WG	Data Science Strategy Working Group
EDRs	electronic dental records
EEG	electroencephalogram
eHOMD	expanded Human Oral Microbiome Database
EHRs	electronic health records
ENCODE	Encyclopedia of DNA Elements
EPA	Environmental Protection Agency
FAIR	findable, accessible, interoperable, reusable
FASTA	text-based format for representing nucleotide or amino acid sequences
FASTQ	text-based format for storing sequence and quality scores
FDA	Food and Drug Administration
FHIR	Fast Healthcare Interoperability Resources
FORCE11	Future of Research Communications and e-Scholarship
GDPR	General Data Protection Regulation
GEO	Gene Expression Omnibus
GIS	Geographic Information System
gnomAD	Genome Aggregation Database
GO	Gene Ontology
GSR	Genomic Summary Results
GTeX	Genotype-Tissue Expression project
GWAS	genome-wide association study
H&E	hematoxylin and eosin stain
HCPCS	Healthcare Common Procedure Coding System
HEAL	Helping to End Addiction Long-term
HIPAA	Health Insurance Portability and Accountability Act
HOMD	Human Oral Microbiome Database
HPO	Human Phenotype Ontology
i2b2	Informatics for Integrating Biology & the Bedside
ICD	International Classification of Diseases
ICD-O	International Classification of Diseases for Oncology
IDC	National Cancer Institute Imaging

	Data Commons
IDS	indigenous data sovereignty
IGVF	Impact of Genomic Variation of Function
IRB	institutional review board
JSON	JavaScript Object Notation
LCM	laser capture microdissection
LinkML	Linked Data Modeling Language
LOINC	Logical Observation Identifiers Names and Code
mCODE	Minimal Common Data Elements
MEPS	Medical Expenditure Panel Survey
MGI	Mouse Genome Informatics
MIAME	minimum information about a microarray experiment
micro-CT	micro-computed tomography
ML	machine learning
MPO	Mammalian Phenotype Ontology
MRI	magnetic resonance imaging
MTA	material transfer agreement
N3C	National COVID Cohort Collaborative
NADCRC	National Advisory Dental and Craniofacial Research Council
NCDB	National Cancer Database
NCI	National Cancer Institute
NCIt	National Cancer Institute Thesaurus
NDA	NIMH Data Archive
NHANES	National Health and Nutrition Examination Survey
NHIS	National Health Interview Survey
NIDCR	National Institute of Dental and Craniofacial Research
NIH	National Institutes of Health
NLM	National Library of Medicine
NLP	natural language processing
NOHSS	National Oral Health Surveillance System
OBO	Open Biological and Biomedical Ontology
OMOP	Observational Medical Outcomes Partnership
OPT	optical projection tomography
OWL	Web Ontology Language
panorex	panoramic x-ray
PATH	Population Assessment of Tobacco and Health
PBRN	practice-based research network
PCORnet	National Patient-Centered Clinical Research Network
PFB	Portable Format for Bioinformatics
PhenX	consensus measures for Phenotypes and eXposures
PRESTO	Pregnancy Study Online
QTL	quantitative trait locus
RDA	Research Data Alliance
RDF	Resource Description Framework
RFI	Request for Information
ScHARe	Science Collaborative for Health disparities and Artificial intelligence bias REduction
SEM	scanning electron microscopy
Smart-DOC	Smart Dental Oral and Craniofacial
SNODENT	Systematized Nomenclature of Dentistry
SNOMED	Systematized Nomenclature Of Medicine
SNP	single nucleotide polymorphism
SRA	Sequence Read Archive
T0/1/2/3/4	stages of the translational spectrum
TCGA	The Cancer Genome Atlas
TCIA	The Cancer Imaging Archive
TEM	transmission electron microscopy
TRUST	Transparency, Responsibility, User focus, Sustainability, Technology
TSV	tab-separated value
UCLA	University of California, Los Angeles
UCSC	University of California, Santa Cruz
VA	Department of Veterans Affairs
VCF	Variant Call Format
WES	whole exome sequencing
WGS	whole genome sequencing