Tissue-to-Bytes: A Catalytic Digital Twin Platform for Consortium-Scale Integration of Single-Cell Omics Data Across the BRAIN Initiative Cell Atlas Network

Shiqiang Tao[1,4], Xiaojin Li[1,4], Yan Huang[1,4], Rashmie Abeysinghe[1,4], Ling Tong[1,4], Shiwei Lin[1,4], Wei-Chun Chou[1,4], Erge Edgu-Fry[1,4], Kimberly A. Smith[2], Lydia Ng[2], Licong Cui[3,4], Guo-Qiang Zhang[1,3,4]*

[1]Department of Neurology, The University of Texas Health Science Center at Houston, Houston, Texas, 77030, USA

[2]Allen Institute for Brain Science, Seattle, WA 98109, USA

[3] McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, 77030, USA

[4]Texas Institute for Restorative Neurotechnologies, The University of Texas Health Science Center at Houston, Houston, Texas, 77030, USA

*Corresponding author: Guo-Qiang.Zhang@uth.tmc.edu, The University of Texas Health Science Center at Houston

ABSTRACT

The NIH BRAIN Initiative Cell Atlas Network (BICAN) is a collaborative effort among neuroscientists, computational biologists, and data scientists to create reference brain cell atlases to be used by the research community as a molecular and anatomical foundational framework for the study of human brain function and disorders. Coordinated sharing and management of the scarce human brain tissue, across lifespans and among the eleven participating BICAN centers and laboratories with anatomical precision, is a formidable undertaking. Creating and operating the necessary data production pipelines with de novo metadata standards prospectively established to adhere to the highest possible levels of Findable, Accessible, Interoperable, and Reusable (FAIR) data principles, present unparalleled challenges that place BICAN into uncharted territories. A digital twin paradigm has been formulated and operationalized to manage advanced single-cell molecular techniques and complexity of data production workflows that are used in the BICAN consortium. The Neuroanatomy-anchored Information Management Platform (NIMP) implements the digital twin paradigm from conception to design, to production. NIMP is an agile, extensible, catalytic infrastructure for integrative and collaborative BICAN consortium-scale FAIR data generation. Digital-twin thinking ensures that NIMP provides tissue-to-bytes, end-to-end integration of data production pipelines spanning brain banks, laboratories, sequencing centers, and data archives. NIMP's digital twin characteristics are manifested in the coupling of real-world workflows occurring in the BICAN laboratories with their digital counterparts moving progressively through verifiable provenance lineage paths. The implementation of NIMP' digital twin features are grounded on advanced informatics strategies that include codification of anatomical contexts; standardized tissue request and sharing workflows, library preparation, sequencing, and single-cell omics data generation and deposition. These workflows and processes are governed by standardized Common Data Elements (CDEs); tracked by blockchain-inspired resource-identifiers (ID); ingested and edited using role-based access control (RBAC); and accessed by query-embedded dashboarding and atlas-enabled interactive resource exploration with dynamically rendered Sankey diagrams. In three years, NIMP has managed data and metadata for 762 human donors, 8,449 brain slabs, and 21,781 sequencing libraries across 11 laboratories, resulting in an estimated total of over 18 million cells for Basal

Ganglia alone. The NIMP digital twin paradigm not only serves BICAN; it provides a blueprint for future-proof data coordination at consortium scale.

Keywords: BICAN · Digital twin paradigm · FAIR data principles · Neuroanatomy-Anchored Information Management Platform (NIMP) · Consortium-scale data coordination · Metadata standardization · Provenance tracking · Single-cell omics · Brain cell atlas · Data interoperability

INTRODUCTION

The NIH BRAIN Initiative Cell Atlas Network (BICAN[1]; RRID:SCR_022794) is a large-scale, collaborative endeavor in the United States that involves neuroscientists, computational biologists, and data scientists aimed at generating comprehensive, multimodal reference atlases of brain cell types across the lifespan of humans, non-human primates, and mice. The project also seeks to elucidate intercellular interactions that contribute to a range of brain disorders. A central mission of BICAN is to develop an open-access compendium of data and tools that will enable the research community to deepen and elevate its understanding of brain structure and function, thereby accelerating advancement in neuroscience. The atlases developed are intended to serve as foundational molecular and anatomical resources for the broader research community, facilitating the study of brain function and disorders. Building upon the findings of the BRAIN Initiative Cell Census Network (BICCN[2]; RRID:SCR_015820), BICAN advances this effort by extending cell and circuit mapping across multiple species, with an emphasis on the human brain.[3,4]

Human brain tissue that qualifies as "neurotypical" is an inherently scarce and valuable resource, typically housed within specialized brain banks that serve as initial custodians. Participating laboratories within BICAN need to be able to identify, review, and request tissue samples to achieve consortium's overall objectives through a joint sampling plan. In some instances, participating laboratories may also seek to further share tissue amongst themselves. Such activities must be performed in a transparent, coordinated manner to ensure that valuable tissue resource is utilized effectively with high anatomical resolution. This necessitates standardized mechanisms that expose available tissue inventories across brain banks, ensuring that laboratories are aware of what types of resources they could potentially request and allowing brain banks to plan and share appropriate tissue blocks accordingly. To further enhance traceability and anatomic precision, all tissue samples are indexed within a Common Coordinate Framework (CCF; RRID:SCR_017764) and annotated with structure labels conforming to ontological nomenclatures, ensuring consistent labeling and tracking of each sample's anatomical origin. Coordinating such activities across more than two dozen BICAN centers and laboratories is a formidable undertaking that underscores the need for a new kind of informatics infrastructure capable of live provenance-tracking from tissue to bytes in this consortium-scale collaboration.

Within BICAN, data production pipelines and metadata standards are developed de novo, prospectively, alongside ongoing scientific activities through Working Groups[5]. To ensure reproducibility and regulatory compliance and to facilitate data access during data generation, these standards need to be constructed in alignment with the highest levels of the Findable, Accessible, Interoperable, and Reusable (FAIR[6]) data principles. These principles define the minimum requirements for data sets that can be consistently located, understood, integrated, and reused across research communities. Adhering to FAIRness during data production is imperative for transparent early-stage data access, which is critical in collaborative, multi-institutional environments such as BICAN. After data are publicly released, FAIRness ensures

that complex, multimodal experimental outputs remain discoverable and reusable throughout the lifetime of the project and beyond BICAN.

Prospective implementation of FAIR data production pipelines and metadata standards requires close coordination among BICAN stakeholders to build consensus across the consortium. Such standards require continuous refinement based on stakeholder feedback during data production, enabling them to remain adaptable to an evolving and expanding set of requirements and emerging new molecular techniques. Such ambitious undertakings and requirements place BICAN's central data infrastructure in uncharted territory and present unprecedented technology and coordination challenges at consortium-scale. Taming these challenges is a monumental effort that demands cohesive strategy, scalable infrastructure and sustained collaboration across all participating institutions; it calls for a new conceptual framework for data coordination.

We address these consortium-scale data coordination challenges through a novel digital twin (DT) conceptual framework. In its most general form, a digital twin is a computerized representation of a physical system that is virtually and dynamically linked to its real-world counterpart, enabling ongoing synchronization, monitoring, and informed decision-making. DTs rely on persistently updated data streams, standardized representations, and bidirectional feedback (i.e., feed-forward and feed-backward) mechanisms that allow the digital model to reflect or even to guide physical operations. The digital twin paradigm has its roots in aerospace and advanced manufacturing as virtual counterparts to physical systems that incorporate real-time bidirectional feedback for monitoring, optimization, and control[7]. Its adoption in biomedicine has expanded to living systems characterized by uncertainty, multiscale dynamics, with stringent ethical constraints[8]. Biomedical DTs integrate clinical, imaging, sensor, and omics modalities into continuously updated computational representations that support prediction, hypothesis testing, and data-driven decision making. The expansion of DTs in biomedicine finds alignment with the maturation of FAIR data infrastructures, cloud-edge computing, artificial intelligence and machine learning methods. Large-scale collaborative consortia have catalyzed this shift by investing in interoperable platforms, governance agreements, and reusable workflows that operate across institutions while maintaining compliance with GDPR, HIPAA, and national policies.[9-15]

Here we employ the DT paradigm to the domain of large-scale consortia data coordination, where data flow and information exchange occur on extended, asynchronous timescales rather than continuously as in traditional DT systems. This approach is well-suited to addressing the formidable challenges and uncharted complexities facing BICAN. The design and full operational realization of the DT paradigm is embodied in our Neuroanatomy-anchored Information Management Platform (NIMP, RRID:SCR_024684) that has been developed under the NIH BRAIN Initiative (U24MH130988) as part of the Coordinating Unit for Biostatistics, Informatics, and Engagement (CUBIE) for BICAN.

Within NIMP, the DT paradigm is implemented through a dynamic coupling between BICAN's physical activities and their corresponding digital footprints. Real-world workflows, spanning both experimental and operational activities executed across BICAN laboratories and centers, are concurrently represented within NIMP. The standardized tissue-to-bytes process for BICAN's data production allows NIMP to accurately mirror and trace resource provenance and data lineage, allowing live reproducible verifications. Because BICAN's requirements emerge in a highly dynamic manner as project unfolds, the classical "waterfall" paradigm for software engineering is unlikely to be effective.[16] Therefore, NIMP adopts an agile software development practice, using diagram-centric design with interface-driven rapid prototyping,

thereby enabling quick adaptation based on evolving requirements guided by the DT design strategy. The platform is extensible and future-proof, allowing new task modules to be integrated seamlessly based on unfolding and modifiable requirements. The DT paradigm positions NIMP as a catalytic force for BICAN: it provides a virtual operations command center for orchestrating key milestone-driven activities for its Working Groups and the development and refinement of Common Data Elements (CDEs) as a part of the metadata standards for BICAN. The NIMP-driven CDE requirements that have originated from NIMP usage and have been vetted through Working Groups ensure adequacy and interoperability in consortium components that include brain banks, sequencing centers, and data archives.

The DT-based conceptual framework and guiding principle are embodied in NIMP's system architecture, which encodes anatomical contexts as a set of core model-cards, ensuring remaining workflows steps to be grounded in precisely defined neuroanatomical structures while supporting diverse molecular modalities. The DT-guided approach requires that every biological entity managed through BICAN is represented consistently and uniquely, harmonized through CDEs, captured as model-cards, and linked to external ontologies for interoperability[17]. For ultimate transparency and reproducibility, NIMP incorporates blockchain-inspired provenance tracking through randomized N-gram Hashing Identifiers (NHash Identifier: RRID:SCR_025313) to ensure tamper-resistant lineage records throughout workflows and data production pipelines[18]. Each NHash ID is dynamically generated at the point of resource entity inception and serves as a globally unique identifier, with an economical, resource-type dependent length allocation. The NHash ID allows every resource entity within the BICAN ecosystem to be unambiguously traced, regardless of local naming conventions or lab-specific identifiers. Role-based access control (RBAC) provides a fine-grained framework for managing NIMP's multi-stakeholder access privileges, allowing differential permission configurations for investigators, brain bank operators, sequencing centers, and coordinating units. Each role is also associated with role-specific query-embedded dashboarding (QED) and faceted browsing to provide convenient access entry points for both technical and non-technical users. The RBAC framework enforces multi-layered validation that verifies user credentials against data operations, ensures metadata completeness against standardized schemas, prevents duplicate entries, and validates anatomical and ontological term usage accuracy before committing resources to the NIMP digital twin. Since digital twins need to provide continuous synchronization between physical and computational states, NIMP exposes robust Application Program Interface (API) layers that enable real-time exchange of identifiers, metadata, and workflow updates with BCIAN stakeholders that operate external laboratory resources and independent data management systems. These API interfaces form the backbone of cross-platform interoperability within BICAN.

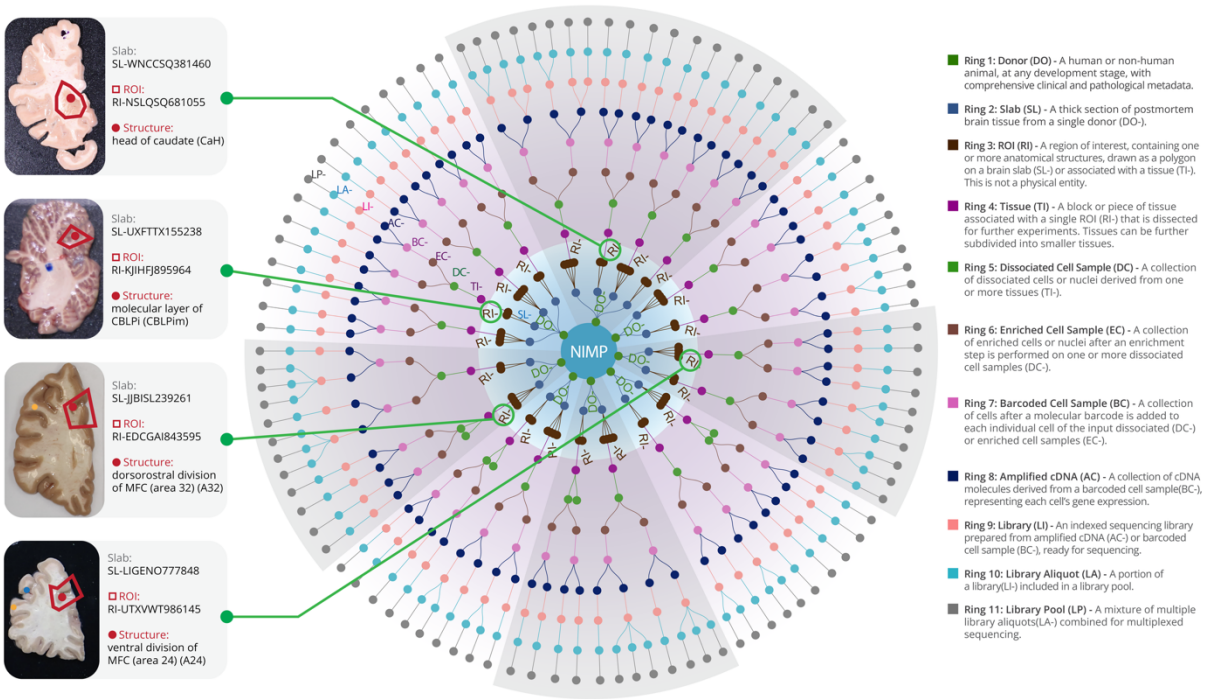

Figure 1. Sample resource linkage rendering using NIMP, capturing how anatomical brain sections are digitally encoded and carried forward through all data production workflow layers to sequencing library pools, enabling tissue-to-bytes integration in a digital twin paradigm.

NIMP provides a common digital twin framework that connects every experimental and computational entity involved in BICAN operations from source to target. It establishes end-to-end lineages from donor and anatomical dissection through tissue processing, molecular profiling, and library generation, enabling bidirectional traceability, automated quality validation, and context-preserving data integration across the entire workflow. Through standardized metadata schemas (CDEs), persistent unique identifiers (NHash), and an interconnected suite of portals and analytical dashboards (see Methods), NIMP ensures that each transformation, derivation, and analytical step is precisely documented, version-controlled, and computationally reproducible. **Figure 1** provides a visual rendering of this digital twin landscape, depicting the progression from specimens to data as a series of color-coded concentric rings spanning donor, slab, region of interest, tissue, dissociated and enriched cell preparations, barcoded cell samples, and finally amplified cDNA and sequencing libraries. These layers represent the pivotal resource states that NIMP manages within a single, coherent digital environment, adhering to BICAN CDEs and metadata standards. By establishing this comprehensive end-to-end digital continuity across neuroanatomical, molecular, and analytical domains, NIMP forms a living digital twin for BICAN's data production ecosystem. The DT paradigm allows BICAN to crystalize and resolve long-standing challenges in data provenance, interoperability, transparency, and multi-modal integration at consortium-scale. NIMP not only makes data findable, accessible, interoperable, and reusable during different stages of data production, but also embeds machine-readable metadata, auto-validation checkpoints, interactive provenance tracking, and bidirectional feedback between physical and digital resources for all stakeholders. The platform implements comprehensive validation mechanisms including role-based verification, metadata schema compliance checks, duplicate detection, and ontological term accuracy confirmation to ensure data quality and consistency across all workflow stages. These features extend FAIR principles into an automated, self-auditing

infrastructure that enhances the overall quality and trust in data and supports sustainable, automation-ready neuroscience data-production at unprecedented depth and scale.

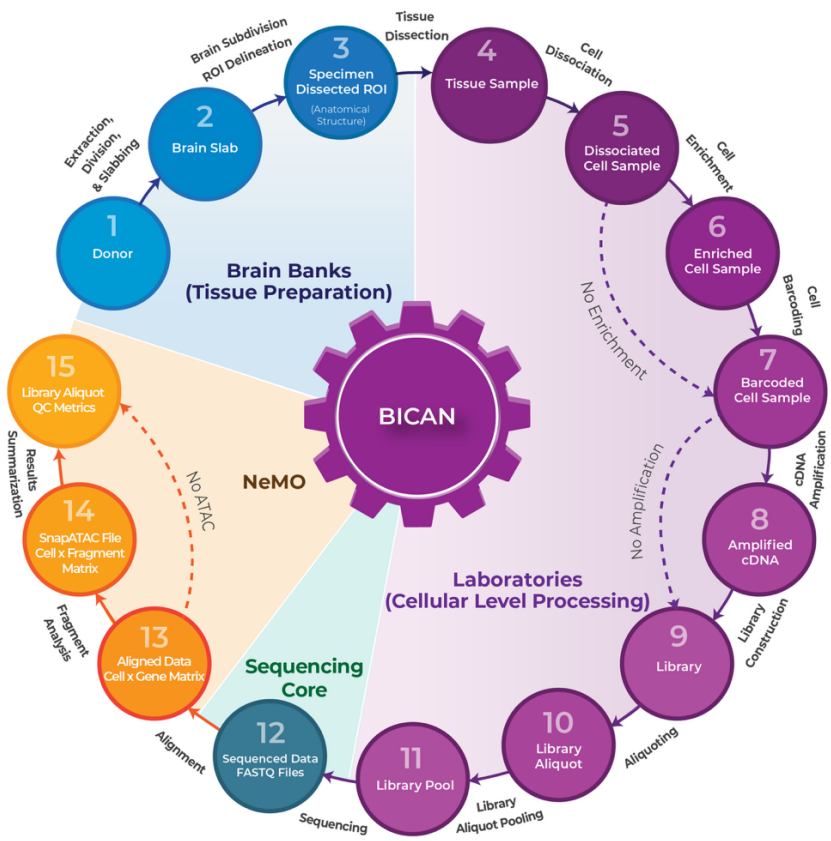

Figure 2. Data workflow among BICAN stakeholders: Brain Banks, Laboratories, Sequencing Cores, and NeMO archive

The DT paradigm virtualizes and orchestrates the entire BICAN data production workflow by maintaining a synchronized correspondence between physical specimens, laboratory actions, and their computational representations. This end-to-end continuum spans brain tissue acquisition, multimodal single-cell processing, sequencing, and archival deposition, with each biological transformation mirrored by a model-carded digital event. To visualize this digital twin conceptual correspondence, **Figure 2** depicts the 14-step standardized workflow linking donor acquisition to reference-grade data outputs. The process begins within brain banks, where donor registration, brain slab preparation, and specimen's region-of-interest (ROI) dissections (steps 1-3) establish the neuroanatomical and provenance foundation for each downstream sample utilization. At this point, the DT is initialized with each physical donor, slab, and ROI, all receiving a respective persistent digital identifier and metadata record that anchors the corresponding real-world coupling.

The workflow then transitions to contributing research laboratories, where steps 4 to11 trace the progressive refinement from tissue to barcoded sequencing library. Tissue samples (step 4) are dissociated into single-cell suspensions (step 5), enriched for target populations (step 6), and processed through molecular barcode incorporation (step 7). Subsequent molecular steps

including cDNA amplification (step 8), library generation (step 9), library aliquoting (step 10), and pooled sequencing libraries (step 11) complete the wet-lab preparation pipeline. Throughout these stages, the digital twin continuously incorporates standardized metadata, quality control (QC) metrics, protocols, and reagent/assay provenance, ensuring that every physical transformation of the biological material is synchronously reflected in its computational counterpart. Step 12 marks the transition to sequencing cores, where pooled libraries are sequenced to generate FASTQ files that anchor the raw molecular output. These sequencing events are automatically reconciled to their upstream samples within the NIMP digital twin, preserving lineage and resource integrity. Finally, the Neuroscience Multi-omic Archive (NeMO, RRID:SCR 016152), the designated BICAN data archive, executes steps 13 and 14, aligning reads to produce molecularly annotated data objects and generating reference-grade outputs including aligned read outputs (BAM files), cell-by-gene matrices, and cell-by-fragment matrices. These downstream products are programmatically linked back through the NIMP digital-twin chain to their physical donor and anatomical origins, completing a closed-loop system in which biological and DT states evolve in lockstep. Together, this workflow demonstrates how the digital-twin paradigm enforces continuous metadata registration, uniform standards, and traceable resource lineage across distributed stakeholders and component infrastructure of BICAN. By coupling each physical transaction with an equivalent digital transaction, the platform ensures that tissue flow, library construction, sequencing, and data archival activities are unified into a transparent, queryable, and quality-controlled production ecosystem, laying the operational DT backbone for consortium-scale, beyond-FAIR human cell-atlas generation.

The digital command center of NIMP covers three tightly coupled portals: (1) the Specimen Portal manages donor and slab data, as well as requests for regions of interest from brain slabs; (2) the Sequence Library Portal oversees the sequencing workflow, enabling the entry of data for components of the BICAN sequencing pipeline including tissues, cell samples, amplified cDNA, libraries, and library pools also supporting the deposition of assay-dependent datasets into modality-specific archives such as NeMO and (3) the Analytics Portal enables users to explore high-level statistical overviews of the growing set of resources generated by BICAN through multi-faceted interrogation of their corresponding digital footprints.

In addition to standards-enforcement and provenance-tracking functionalities, NIMP offers a suite of tools for data visualization and interrogation to address the entire scope of consortium needs by anatomical origins, by resource identifiers, or by resource types and laboratory-specific local identifiers. Sankey diagrams rendered live support interactive provenance tracking by allowing users to follow data flow pathways from donor to tissue to library[19]. APIs and data export interfaces allow NIMP to interoperate seamlessly with external repositories and computational environments, whereby enabling scalable data ingestion and export, supporting the enforcement of quality control policies, and ensuring reproducible analytical results using Terra-based single-cell workflows[20].

Over its first three years of operation, NIMP has generated and accumulated metadata from 762 human donors, 8,449 brain slabs, and 21,781 sequencing libraries across 11 participating laboratories, collectively encompassing an estimated 18,341,279 cells within the Basal Ganglia[21] alone. Beyond supporting BICAN objectives, the NIMP DT endeavor pioneers a promising blueprint for next-generation consortium-scale data coordination and management.

RESULTS

NIMP transforms the conventional one-way data-capture model into a bidirectional digital-twin feedback system, in which laboratory actions continuously update their digital representations, and digital analytics in turn inform experimental decisions in real-time. Within this framework, every specimen, library, and dataset exists simultaneously in both physical and digital form: as tissues and libraries in the laboratory pipeline, and as CDE model-carded data resources within NIMP's metadata ecosystem. When a user performs an operation in the real world such as dissecting a brain region or pooling sequencing libraries, its digital twin is created or updated in NIMP through web interfaces or standardized APIs. Conversely, digital validations, quality metrics, and visual analytics automated within NIMP propagate back to dashboards that inform subsequent physical workflows, such as prioritizing tissue usage, refining ROI selection, or adjusting sequencing strategies. This two-way coupling establishes a closed feedback-loop between laboratory operations and data coordination command-and-control, effectively allowing NIMP to synchronize with laboratory activity states. BICAN leverages the NIMP DT to identify workflow bottlenecks, predict resource needs, and signal quality issues before they propagate further downstream. The following subsections detail how each of NIMP's three operational portals: the Specimen Portal**,** Sequencing Library Portal, and Analytics Portal, embodies this bidirectional feedback mechanism during each stage of the end-to-end tissue-to-bytes data production pipeline.

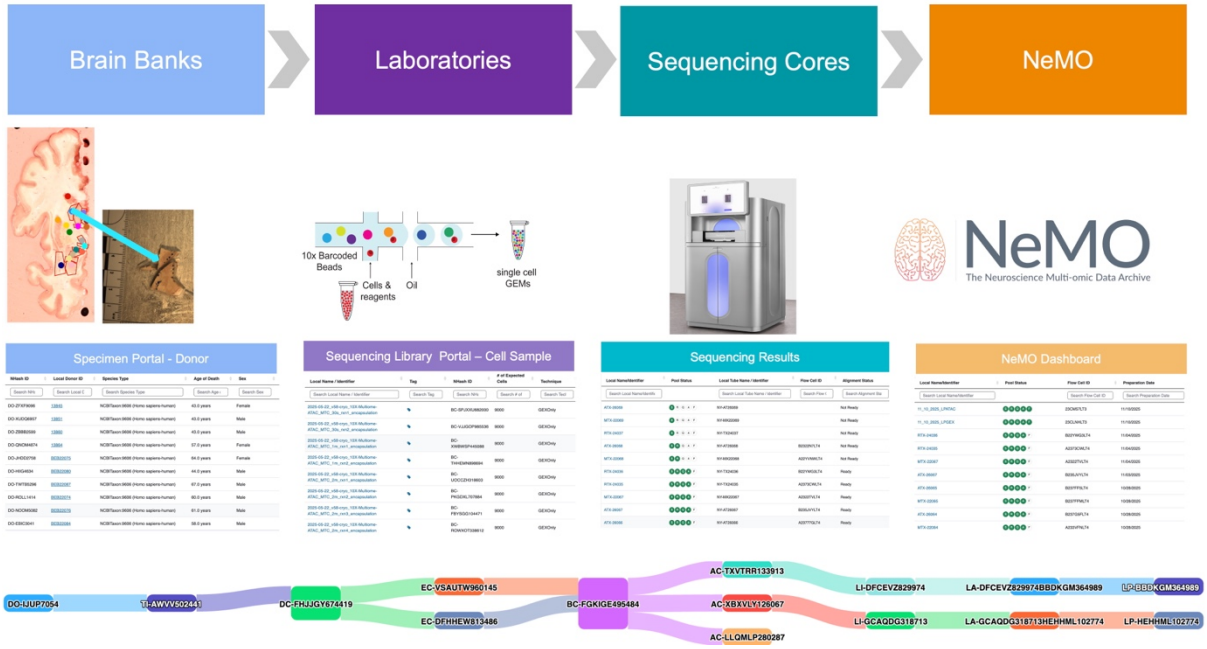

Figure 3. NIMP implements an end-to-end integration of BICAN data production following the digital twin paradigm

Feed-forward: Specimen Portal

To support consortium-scale coordination between brain banks and laboratories, we developed the Specimen Portal as the feed-forward component of our bidirectional digital-twin workflow. In this paradigm, the portal operationalizes the specimen-request pipeline by moving information and decisions downstream: from donor registration to brain slab cataloging, to ROI annotation during tissue requests and request approvals. Later stages of the workflow return data resources in NIMP that feedback into the digital twin for command-and-control by stakeholders. The complete workflow is summarized in **Figure S1**. The pipeline begins with

donor registration, where brain banks enter standardized anatomical and clinical metadata for each donor. These metadata include general demographic information, brain specimen descriptors, neuro- and non-neuropathological diagnoses, medical and family history, infectious disease screening, and toxicology results, all parts of the BICAN CDEs. After donor-level metadata are captured, brain banks upload brain slab collections and detailed annotations for each individual brain slab, establishing the structural inventory from which ROIs can be annotated and requested.

Laboratory investigators initiate the next stage by submitting ROI requests through the Specimen Portal. Requests specify donor identity (via de-identified BICAN accession IDs), slab identifiers, and anatomical structures to be dissected for experimental profiling, all uniquely identifiable using NHash IDs. The portal then routes each request to the corresponding brain bank personnel for review and approval, providing a unified interface for evaluating donor suitability, availability of remaining tissue, and alignment with institutional and consortium tissue and data access policies. Upon approval, brain bank personnel use NIMP to annotate and mark dissected regions directly on digital brain slab representations, thereby updating the real-time tissue inventory and creating a fresh record of what material has been allocated, used, or remains available. These actions performed at the physical tissue level are captured digitally and propagated forward as model-carded transactions within the NIMP DT, enabling provenance tracking and downstream linkage to laboratory-generated single-cell libraries and sequencing outputs.

The Specimen Portal provides an integrated mechanism for standardized tissue request management**,** traceable ROI dissection, high anatomical resolution, and dynamic inventory tracking**,** forming the feed-forward arm of the digital-twin paradigm that ensures that upstream tissue operations are consistently aligned with the data production and archival processes that follow.

Feed-forward: Tissue Dissection with Region of Interest (ROI) Annotation

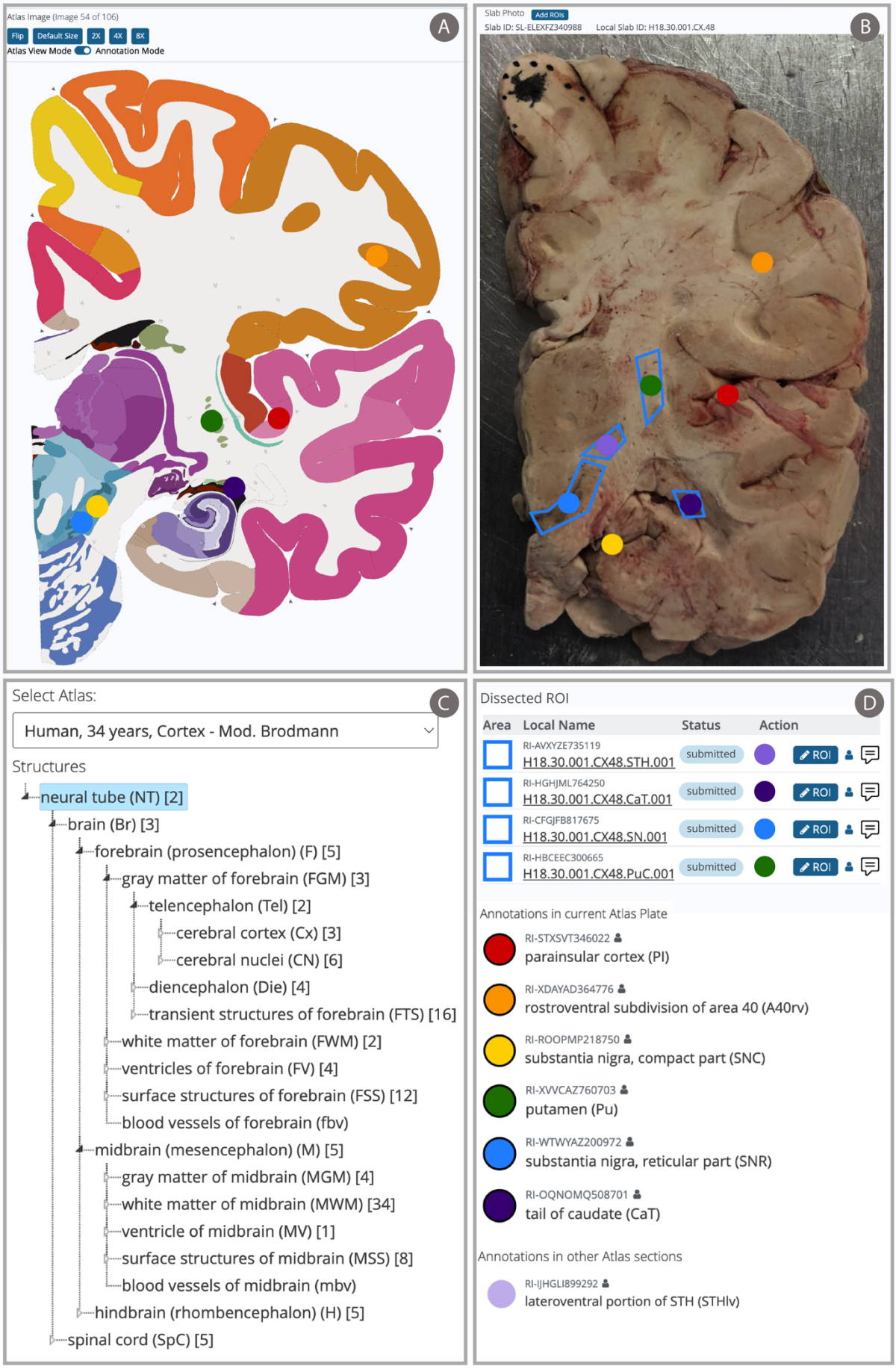
A
Atlas Image (Image 54 of 106)
Flip
Default Size
2X
4X
8X
Atlas View Mode
Annotation Mode
B
Slab Photo
Add ROIs
Slab ID: SL-ELEXFZ340988
Local Slab ID: H18.30.001.CX.48
C
Select Atlas:
Human, 34 years, Cortex - Mod. Brodmann
Structures
neural tube (NT) [2]
brain (Br) [3]
forebrain (prosencephalon) (F) [5]
gray matter of forebrain (FGM) [3]
telencephalon (Tel) [2]
cerebral cortex (Cx) [3]
cerebral nuclei (CN) [6]
diencephalon (Die) [4]
transient structures of forebrain (FTS) [16]
white matter of forebrain (FWM) [2]
ventricles of forebrain (FV) [4]
surface structures of forebrain (FSS) [12]
blood vessels of forebrain (fbv)
midbrain (mesencephalon) (M) [5]
gray matter of midbrain (MGM) [4]
white matter of midbrain (MWM) [34]
ventricle of midbrain (MV) [1]
surface structures of midbrain (MSS) [8]
blood vessels of midbrain (mbv)
hindbrain (rhombencephalon) (H) [5]
spinal cord (SpC) [5]
D
Dissected ROI
Area
Local Name
Status
Action
RI-AVXYZE735119
H18.30.001.CX48.STH.001
submitted
ROI
RI-HGHJML764250
H18.30.001.CX48.CaT.001
submitted
ROI
RI-CFGJFB817675
H18.30.001.CX48.SN.001
submitted
ROI
RI-HBCEEC300665
H18.30.001.CX48.PuC.001
submitted
ROI
Annotations in current Atlas Plate
RI-STXSVT346022
parainsular cortex (PI)
RI-XDAYAD364776
rostroventral subdivision of area 40 (A40rv)
RI-ROOPMP218750
substantia nigra, compact part (SNC)
RI-XVVCAZ760703
putamen (Pu)
RI-WTWYAZ200972
substantia nigra, reticular part (SNR)
RI-OQNOMQ508701
tail of caudate (CaT)
Annotations in other Atlas sections
RI-IJHGLI899292
lateroventral portion of STH (STHlv)

Figure 4. ROI annotation interface in the Specimen Portal. (A) Atlas ontology browser for standardized anatomical term selection; (B) Pinpoint tool for landmark registration to brain atlas sections; (C) Freehand polygon drawing interface for delineating precise ROIs on brain slab images; (D) Information panel displaying all ROIs and pins with associated metadata

The ROI annotation module is an important digital twin component within NIMP, faithfully linking physical brain tissue specimens with their digital neuroanatomical representations. The web interface (**Figure 4**), built on the Annotorious.js library, allows investigators to delineate precise, structure-labeled ROIs directly onto high-resolution brain slab images by freehand polygon drawing, auto-referencing the Allen Human Brain Atlas (RRID:SCR_017764) for anatomical structure labeling, and activating the Specimen Portal CDE model-card for tissue request and sharing within the BICAN consortium.

Unlike alternative strategies, this module precisely mirrors physical procedures performed on brain slabs, documenting all anatomical dissections with standardized brain structure terms and spatial CCFs. Each ROI delineation captures geometric boundaries and embeds semantic context anchors using Allen's Human Brain reference atlase, creating a standardized, reproducible, and auditable digital copy of tissue model-card, properly stored in NIMP's backend relational databases. To maintain tissue resource integrity, this module performs automated checks for area collisions, boundary closures, and topological consistency before committing the requested transaction in the NIMP digital twin. Advanced tools were used to further support complex tissue-related workflows: multi-step dissection functions subdivide large ROIs into progressively smaller areas, while laminar sectioning tools partition regions into stratified layers essential for depth-dependent molecular profiling. Downstream, the ROI module facilitates transparent tissue resource information sharing, providing receiving laboratories with full spatial and semantic context without manual re-entry of tissue model-card metadata.

The ROI annotation module exemplifies the synergistic and catalytic nature of the NIMP digital twin by transforming idiosyncratic, siloed documentation styles about tissue resources in local brain banks into a common, shared, transparent digital twin platform. It enables transparent request and sharing processes where requesters view which brain slab portions have been allocated, which remain available, and how specimens have been utilized, maximizing effective use of scarce human donor tissue. Real-time visibility into tissue availability prevents redundant sampling, reduces waste, and facilitates coordinated sampling strategies aligned with the BICAN Joint Sampling Plan. This ROI annotation module also establishes auditable experimental trajectories from planning to data generation, ensuring tissue requests and subsequent analyses are anchored to reference neuroanatomy. This catalytic paradigm drives tissue metadata development and standardization across the consortium through a unified annotation strategy, facilitating consistent use of brain structure labels and spatial registration protocols.

Feed-forward: Sequencing Library Portal

In tandem with the Specimen Portal, the Sequencing Library Portal serves as the second major feed-forward component of the digital-twin architecture. While the Specimen Portal manages donor-to-ROI tissue allocation, the Sequencing Library Portal extends the workflow downstream through all wet-lab processing steps, capturing the generation, quality control, and packaging of sequencing-ready libraries (Figure S2).

The Sequencing Library Portal is organized into seven functional modules: Tissue, Cell Sample, Amplified cDNA, Library, Aliquot QC, Library Pool, and Library Package, which map directly onto the standard experimental stages used across BICAN laboratory sites. Each module captures both the provenance and the processing history of its corresponding biological material in standardized model-card forms. The workflow begins when investigators retrieve approved tissue ROIs from the Specimen Portal and register them in the Tissue module. Subsequent steps including dissociation, cell or nuclei enrichment, barcode incorporation, and cDNA amplification, are tracked through the Cell Sample and Amplified cDNA modules. Libraries generated from amplified cDNA or barcoded Cell Sample are recorded in the Library module, followed by quality-controlled aliquots entered in the Library Pool module.

To support downstream sequencing operations, investigators may combine aliquots from different libraries into Library Pools, which are then assembled into Library Packages for submission and shipping to designated sequencing centers. The portal provides a unified interface for registration of model-cards that represent packaging metadata, sequencing requests, sample manifests, and associated QC information, allowing sequencing cores to receive standardized, complete submissions regardless of the originating laboratory.

Together, the Specimen Portal and Sequencing-Library Portal coordinate the multi-center workflow spanning brain banks, library laboratories, sequencing cores, and data archiving systems. The two portals work in tandem to manage essential data-production resources including donors, brain slabs, tissue ROIs, cell samples, libraries, aliquots, pools, and packages, and ensure consistent provenance tracking that encapsulates the full continuum of activities involved in sample processing. At present, 11 laboratories use these portals daily to capture, validate, transmit materials, and share data and status information within BICAN, enabling FAIR, interoperable inputs into the digital twin and onto the downstream NeMO data archive.

Feed-backward: Analytics Portal

To complete the bidirectional digital-twin feed-forward and feed-backward loop, we developed the NIMP Analytics Portal as the primary feed-backward interface that returns harmonized, spatially contextualized, and model-card structured information to BICAN stakeholders. Whereas the Specimen and Sequencing-Library portals capture the forward trajectory of tissue and library processing, the Analytics Portal closes the loop by enabling investigators to interrogate and visualize the data generated across BICAN. Two interactive web interfaces: Data Overview Heatmap and Atlas Browser provide complementary entry points for multi-scale exploration of NIMP's digital twin asset.

Feed-backward: Data Overview Heatmap

The heatmap interface provides an integrated environment for examining BICAN sequencing resources within the brain atlas ontology, enabling users to evaluate how samples are distributed across anatomical structures and contributing laboratories or barcoded cell sample techniques. This comprehensive view facilitates assessment of coverage, identification of underrepresented regions, and detection of redundancies, supporting verification, strategic planning, and informed resource allocation that ultimately informs progress towards BICAN's Joint Sampling Plan. Within the NIMP digital twin architecture, the interface functions as a real-time representation of ongoing data generation, offering a continuously updated perspective that supports coordinated oversight, quality control, and efficient monitoring and management of multi-center contribution towards shared BICAN scientific objectives.

Users can interactively refine the visual rendering scope through the selection of species filters, hierarchical ontology-based menu items, and text-based search, allowing both broad surveying

and precise anatomical targeting. The heatmap dynamically responds to user-defined criteria, enabling iterative visual inspection of sampling patterns and rapid identification of anomalies or inconsistencies. Each matrix cell is linked to a detailed table of the corresponding sequencing records, where users can examine metadata, apply additional filters, and perform batch-export of curated subsets for offline analysis. This design ensures a seamless transition from aggregated trends to record-level data support, allowing investigators to inspect and validate data completeness, investigate suspected gaps, and help debug and reconcile data quality issues by tracing back to their origin of responsible parties.

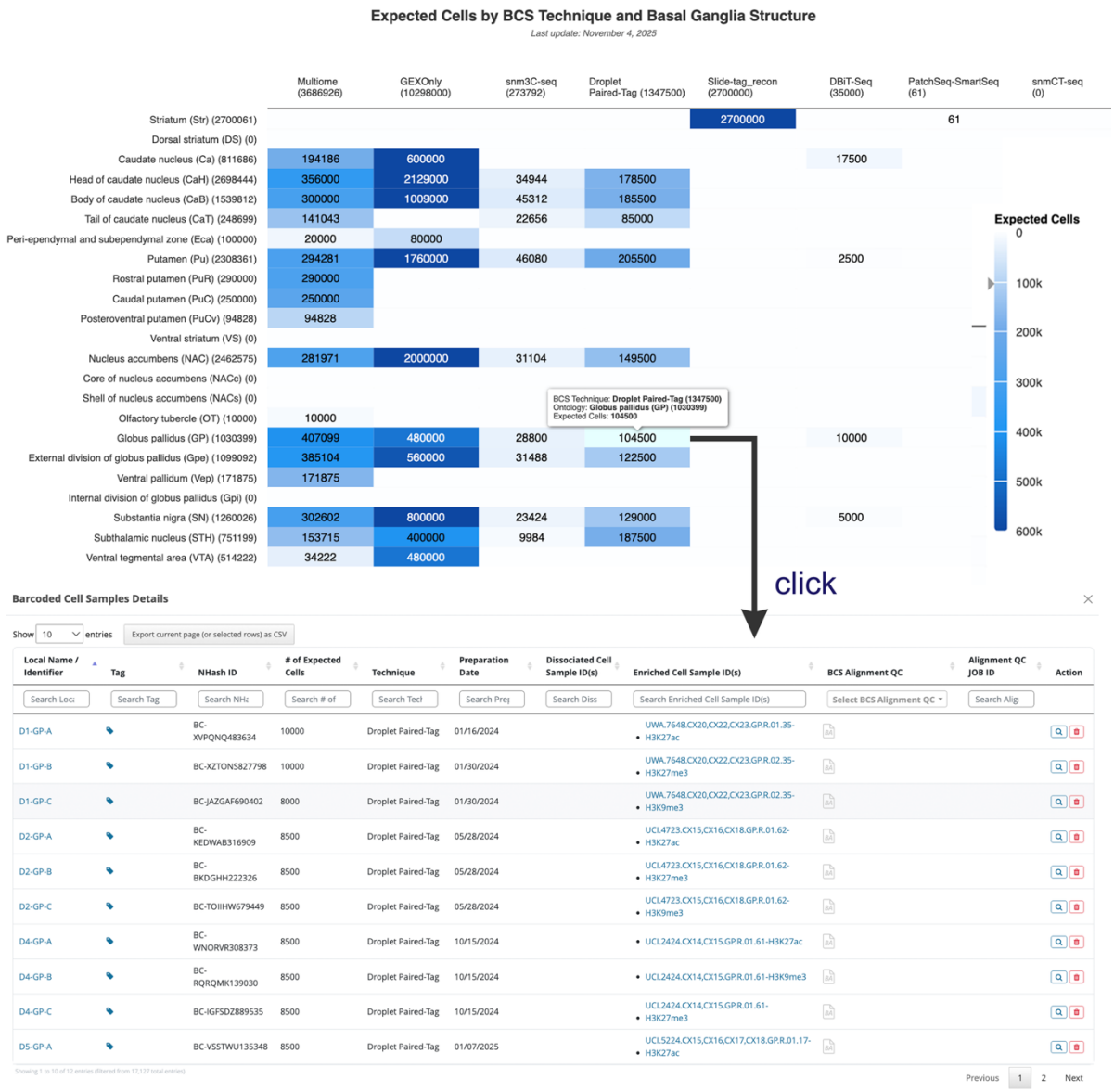

Figure 5. Heatmap interface representing the distribution of sequencing resources across basal ganglia structures, with this region presented as a representative example.

The heatmap interface is shown in Figure 5, using the basal ganglia as an illustrative example. By aligning barcoded cell sample techniques along the x-axis and basal ganglia substructures along the y-axis, the interface provides a real-time overview of sequencing resource distribution. The interactive matrix (Figure 5, top) encodes the number of resources processed by each technique - structure pair, with color intensity indicating relative count-levels.

Selecting a cell reveals the corresponding record-level table (Figure 5, bottom), which supports subsequent filtering and structured text export to facilitate detailed examination of underlying resources offline. The heatmap rendering can be further refined by species, and the integrated brain atlas ontology browser allows a user to limit the view to specific anatomical regions using either hierarchical navigation or text-based search.

Feed-backward: Atlas Browser

The Atlas Browser provides a spatially anchored interface within the NIMP digital twin architecture, linking the physical anatomy of the brain to its digital representations. It enables a user to explore anatomical structures in 2D and 3D while maintaining live connections to specimen metadata and provenance records, supporting multiscale interaction across structural, molecular, and experimental dimensions. This integration allows users to intuitively perform first-line anatomical coverage assessment, visually inspect underrepresented regions, and contextualize molecular or transcriptomic data within precise neuroanatomical locations, thereby informing study priorities, monitoring data-production throughput, and guiding resource allocation. By projecting the brain's physical organization directly within the digital framework, the Atlas Browser reenforces the defining characteristics of the digital twin paradigm, providing a continuously updated footprint of the tissue-to-bytes transformation.

With the 2D Atlas Browser, users navigate atlas plates through interactive sliders, hierarchical structure panels, and text-based search tools, refining their visual focus on specific regions or structural hierarchies. Structure term selection triggers dynamic retrieval of linked specimens, including donor and tissue details, ensuring that exploration and data access remain synchronized. Multiple visualization modes, including color-coded annotations and layered overlays, enhance interpretability and facilitate the comparison of adjacent structures or sequential sections leveraging the capability of human visual perception. The 3D Brain Viewer extends this functionality to a volumetric brain representation, preserving spatial alignment between 2D plates and reconstructed meshes. Users can rotate, zoom, pan, and adjust opacity to examine fine-scale anatomical relationships within the global brain context. Ontology-guided filtering and cross-referencing allow researchers to highlight relevant structures, assess connectivity, and link observations directly to the Resource Provenance Visualization Engine (RPVE[22]), reinforcing the bidirectional interplay between the physical system and its digital twin. In this way, the Atlas Browser provides a continuously updated, interactive, and analytically robust environment, ensuring that the digital representation of the brain accurately reflects the evolving state of BICAN resources and sets a foundation for the broader reproducible, data-driven neuroscience discovery.

The 2D Atlas Browser supports detailed visualization and inspection of individual atlas plates, organized into six interrelated modules as shown in Figure S3 (A-F). The Atlas Dropdown Menu (Figure S3.A) enables selection among multiple atlases from comparative and developmental perspectives. The Whole-Brain Reference panel (Figure S3.B) provides global anatomical context, while the Atlas Plate Navigator (Figure S3.C) allows smooth transitions across plates. The Plate Display (Figure S3.D) renders plates with color-coded annotations, and the Structure-in-Plate panel (Figure S3.E) enumerates annotated structures with counts and spatial feedback. The Cross-Plate Search Module (Figure S3.F) supports structure-based queries across the entire atlas, linking directly to detailed donor and tissue model-card items.

3D Brain Viewer (Figure 6) enables interactive rendering of reconstructed brain views with rotation, zoom, and panning, along with tooltips showing structure-specific metadata (Figure 6.A). Four views demonstrate its versatility: lateral view with all plates and annotations (Figure

6.A1), dorsal view highlighting a subset (Figure 6.A2), inferior view without overlays for clear visualization (Figure 6.A3), and oblique view isolating individual plates (Figure 6.A4). The Control Panel allows adjustment of plate range and opacity (Figure 6.B), and ontology-guided filtering supports targeted selection of anatomical structures (Figure 6.C). Tooltips link to Sankey diagrams (Figure 6.D) tracing relationships between structures and sequencing resources, with access to detailed specimen and model-card tables (Figure 6.E).

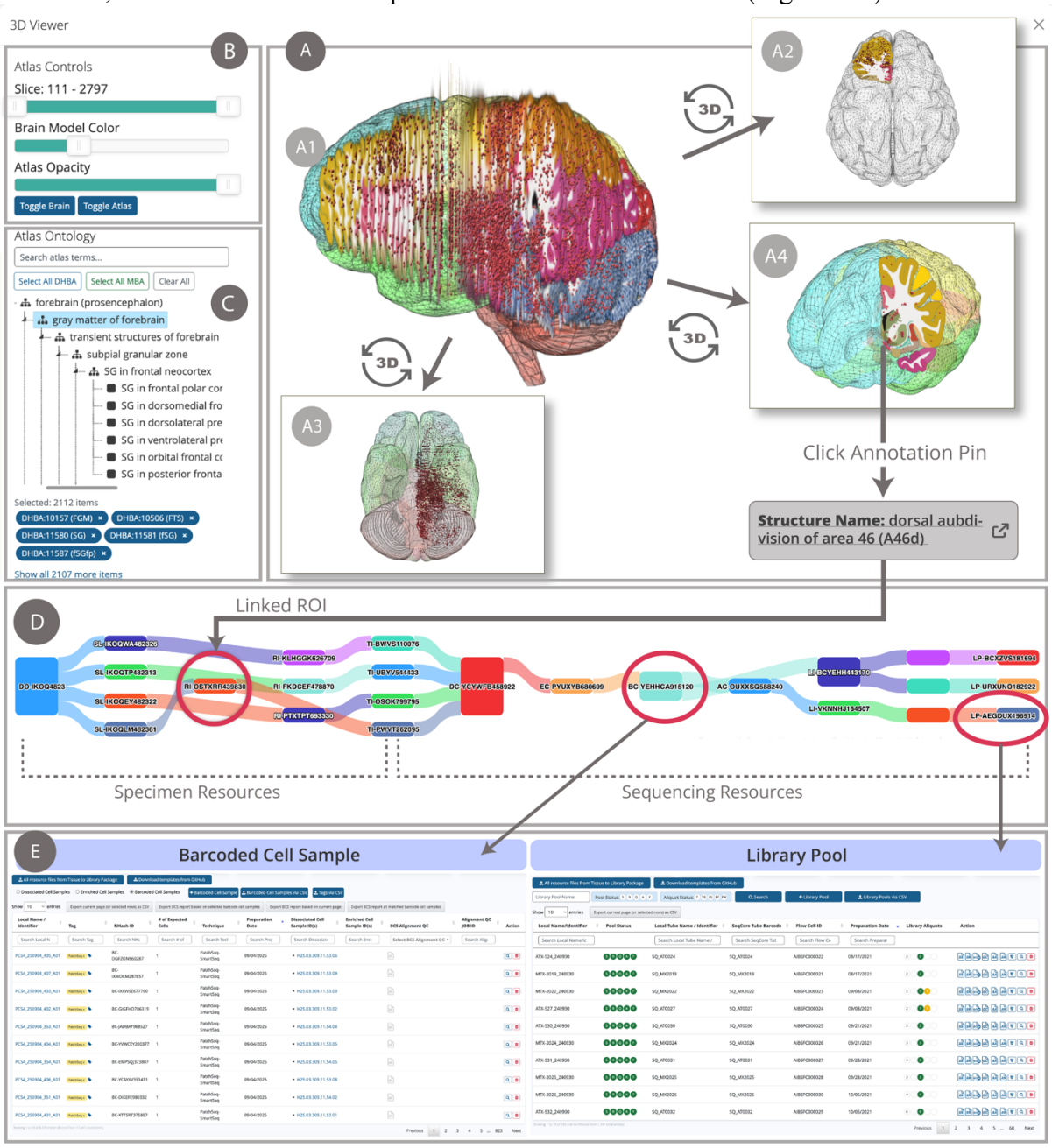

Figure 6. 3D Brain Viewer interface for interactive exploration of reconstructed brain models. (A) Tooltips displaying structure-specific metadata; (A1) Lateral view with all plates and annotations; (A2) Dorsal view highlighting a subset of structures; (A3) Inferior view without overlays; (A4) Oblique view isolating individual plates; (B) Control Panel for adjusting plate range and opacity; (C) Ontology-guided filtering for targeted structure selection; (D) Sankey diagrams linking structures to sequencing resources; (E) Tables providing detailed specimen and dataset information.

DISCUSSION

The adoption, implementation, and operation of the DT paradigm through NIMP have played a catalytic role that helped propel the BICAN data-production ecosystem with essential consortium-scale FAIRness that encompasses the entire tissue-to-bytes pipelines as resources were generated at inception. By mirroring the complete lifecycle of biospecimens from donor registration and tissue dissection to sequencing, analysis, and data release, the NIMP Digital Twin framework ensures that each specimen and derived dataset is linked through persistent, verifiable identifiers. This tissue-to-bytes continuum creates a unified digital representation of biospecimen transformation, allowing researchers to navigate seamlessly between biological entities and their digital footprints.

The DT paradigm fosters synergistic collaboration and catalytic requirement-elicitation across BICAN components. The NIMP DT provide additional meta-level feedback for consortium operation: (a) direct system feedback, which continuously monitors metadata completeness and technical performance, and (b) knowledge-level feedback, which aggregates insights across consortium to identify procedural gaps, protocol bottlenecks, and emerging priorities. This meta-level feedback serves as valuable inputs for optimizing collaboration and coordination, monitoring progress towards milestones, facilitating communication, and sets agenda for Working Groups, whereby supporting operational transparency and consortium integrity.

At the core of the NIMP DT is the NHash identifier algorithm, which provides a robust, collision-resistant mechanism for communicating and tracking resource entities across laboratories, sequencing cores, and data archives in a readable, space-economical fashion rather than one-size fit all (as in UUIDs). Just-in-time generation of new NHash identifiers takes place precisely when new digital items come into existence in NIMP, continuously feeding into live dashboards, Sankey-based visualizations, and API-driven interfaces that provides a uniform, unique, and global handle for all upstream and downstream consortium data-production and release activities.

Crucially, the digital twin design drives metadata standardization across a distributed consortium. DT mandates consistent schemas, controlled vocabularies, and interoperable model-card formats, thus requiring FAIRness as upstream as possible. It not only makes resources Findable, Accessible, Interoperable, and Reusable, but also help achieves a new level of FAIR maturity by putting fit-for-purpose, applicability, and usability at the driver for Working Group activities. NIMP's prospective quality control mechanisms detect and flag deviations during the data entry phrase and ensure reproducibility and reliability data products that must pass the hands of multiple stakeholders.

Beyond the mandate of a common conceptual framework, expandability and refinability are also hallmarks of the NIMP digital twin architecture. The adoption of the latest informatics strategies in implementing NIMP makes it a nature part of the DT in accommodating new requirements as they emerge through all stages of open-ended consortium collaboration. As BICAN continues to evolve, this extensible infrastructure enables continuous refinement of digital twins to integrate new resource types, workflows (such as MERFISH[24]), and archives, in aspiration to parallel the transformative trajectory once achieved by the Human Genome Project but now at single-cell and systems neuroscience resolution.

Feedback from users across BICAN has underscored NIMP's consistent roles for quality control, data release coordination, and process transparency that would not be possible without the DT paradigm. By coupling real-time interactive data and metadata tracking with interactive visual analytics and standardized identifiers, the NIMP digital twin bridges the gap between

biological specimens and digital-resource knowledge, setting a new standard for data stewardship and integrity in large-scale neuroscience collaboration. It universally elevates the level of trust in data and metadata that are produced across the consortium, instilling the same level of confidence in data integrity for laboratories and users that are not directly involved in specific data production, because of a consortium-scale digital twin.

Acknowledgements

This project is funded under the NIH BRAIN Initiative award U24MH130988 as part of the Coordinating Unit for Biostatistics, Informatics, and Engagement for BICAN. We thank Dr. Yong Yao (NIH/NIMH) and Dr. Adi Cymerblit-Sabba (NIH/NINDS) for their continuous feedback and insightful suggestions throughout the development of the NIMP platform. We are grateful to Dr. Michael Hawrylycz for his feedback on the design and interpretation of the analytics-portal heat map and to Rachel Dalley for her expertise in systematic ROI annotation. We acknowledge Arron Garcia for his contributions to the planning and effort that enabled individualized brain slab generation. We thank Suvarna Nadendla for her comments and feedback on the design and implementation of the NIMP - NeMO information-exchange interface. We also thank the sequencing teams at the New York Genome Center (NYGC) and the Broad Institute for their support in implementing and refining the sequencing library portal. Finally, we thank the BICAN Specimens Working Group (Ed Lein, Charles Vanderburg, Kiku Ichihara, Emily Finn), Data Ecosystems Working Group (Carol Thompson, Owen White, Tim Tickle) and the Metadata and Ontologies Integration Working Group (Jim Zheng, Hua Xu, Patric Ray, Maryanne Martone) for their ongoing input, discussions, and critical evaluation, and for Yasmeen Hussain for coordination and communication support, all of which have strengthened the robustness, interoperability, and long-term sustainability of the system.

Author Contributions

Conceptualization: GQZ
Methodology: GQZ, ST, XL, YH, RA
Software design and development: ST, XL, YH, SL, LT, WCC
System Performance: ST, XL, YH
Testing & Enhancement: ST, XL, YH, RA, LT, KAS, GQZ
Data Curation: ST, XL, YH, KAS, LT, SL
Visual Diagrams: ST, XL, YH, WCC, GQZ
Domain Resources: KAS, LN, LC
Writing & Editing: initial outline: ST, EEF; subsequent versions: all authors
Finalizing Review, Editing, Submission: GQZ
Supervision & Project Administration: GQZ, ST, XL
Funding Acquisition: GQZ

*Correspondence: Guo-Qiang Zhang

Declaration of Interests

The authors declare no competing interests.

METHODS

Metadata Standardization and Terminology Management

Since the Digital Twin paradigm depends on a shared conceptual framework to enable consistent tracking of key resource items in the physical world through a standardized digital representation, it is essential that attributes of the resource items are represented with sufficient details for use across BICAN. As BICAN data production relies a diverse set of contributing

resources, Common Data Elements (CDEs) provide the foundation for interoperability, consistency, and reproducibility across the consortium. Within BICAN, CDE standards are developed prospectively to align with ongoing scientific and data production activities. Although NIMP is driven by these CDE standards, it also serves as a primary catalytic platform for their origination, design, curation and subsequent dissemination. CDE standards are developed as BICAN is formed, through its Working Groups, this prospective nature of the development of CDEs requires close coordination among BICAN stakeholders to ensure that standards remain synchronized with the emerging developments across the consortium. Continuous refinement of the CDE standards through NIMP DT played an essential role to reflect evolving needs of the consortium.

For each data element, an NHash ID is assigned, and associated model-card that includes the data element's name, description, data type, permissible values (where applicable), and nullability are implemented and version-controlled and maintained in NIMP. CDEs are released through NIMP using two channels: (1) the Terminology Browser, which provides an interactive interface for exploring the data elements; and (2) a public GitHub repository, which supports version control and offline access without requiring a NIMP login.

NIMP also hosts several reference brain atlases, including the Allen Developing Human Brain Atlas, Mouse Brain Atlas, and Penn-CHOP Human Neonatal Brain Atlas that are used for tissue annotation and tracking. Where possible, cross-reference among these and other anatomical systems are maintained. Additional atlases can be ingested and integrated into NIMP as they become available.

Terminology Browser. The Terminology Browser (Figure S4) provides an integrated interface for exploring and inspecting terminological and metadata resources. It comprises four main tabs: Terminology Table, Tree Structure, Atlas Terminology, and Categorical Variable Code-Label Reference. The Terminology Table tab (Figure S4.A) presents all data elements in a structured format, with an option to export content for downstream analysis. The Tree Structure tab (Figure S4.B) organizes elements hierarchically, enabling navigation through nested structures and inspection of associated value sets, supporting data interpretation. The Atlas Terminology tab (Figure S4.C) allows navigating and browsing of ingested brain atlases, with expandable child nodes, detailed metadata, and a search function for rapid term lookup. The Categorical Variable Code-Label Reference tab (Figure S4.D) displays value sets for categorical variables, showing both codes and labels, with a search functionality to help consistent and standardized semantic interpretation across studies. These features facilitate efficient exploration, retrieval, and contextualization of metadata, enhancing data interoperability and reproducibility within the NIMP digital twin paradigm.

Block-chain Style Provenance-aware Resource Identifier

To support trustworthy, provenance-aware data integration across BICAN, NIMP implements a blockchain-style identifier generation algorithm using randomized N-gram Hashing (NHash). NHash functions as a lightweight, distributed mechanism for generating globally unique, collision-avoidant, and tamper-resistant resource identifiers without resorting to an external centralized issuing authority. By deriving each identifier from a random seed and well-defined input attributes, and by enabling deterministic regeneration for verification, NHash provides an audit-friendly trail of provenance resource lineage that is analogous to cryptographic transaction signatures.

At its core, NHash transforms resource-specific attributes, such as donor ID, project ID, or tissue accession number into short N-gram hashes that cannot be reverse-engineered to reveal

the original upstream metadata fragments being used. These hashes are then combined with a randomly generated seed and optionally encrypted (e.g., via a shift-cipher step) to produce a final identifier whose structure encodes provenance while preserving the metadata fragments used. This two-phase design, consisting of (1) generating an intermediate N-gram string and (2) cryptographically transforming it into the final token, ensures that identifier errors can be detected and invalid identifiers (e.g., formed during manual editing) can be detected at any time by recomputing them from their seed.

Originally developed for a disease research collaborative center, NHash has evolved from a participant-identifier tool into a general algorithm supporting a provenance-tracking engine capable of tracing heterogeneous and mutually dependent research resources, including donors, tissues, cell samples, amplified cDNA, sequencing libraries, library aliquots, and library pools. Its integration into NIMP demonstrates its scalability as a consortium-scale identifier management system, enabling traceable links between upstream and downstream resources while maintaining interoperability and provide tamper-proof measures.

For example, the NHash ID DO-IHNG4035 in Figure S5 encodes four provenance elements: (1) DO designates the resource type (donor); (2) 4035 is the randomly generated seed; (3) IH is the N-gram hash derived from the subject identifier and seed; and (4) NG is the corresponding hash derived from the project identifier and seed. Together, these components form verifiable and provenance-preserving chains that supports robust resource tracking throughout the research workflow, embodying the essential properties of a blockchain-style system - immutability, consistency, and trust without centralized control.

Resource-Provenance Visualization

Large-scale consortia such as BICAN generate complex, multi-stage workflows spanning specimen acquisition, processing, and data generation. Within the NIMP digital twin paradigm, maintaining a precise and continuously synchronized correspondence between physical samples and their digital representations is essential for integrity, reproducibility, and coordinated oversight. This requires a provenance layer that can capture, update, and display the evolving state of materials and digital assets across centers. Conventional tools often fail to remain interpretable as workflows expand in depth and heterogeneity, making it impossible to audit dependencies, identify inconsistencies, or investigate how datasets arise from upstream biological materials. The Resource-Provenance Visualization Engine (RPVE) addresses these challenges by providing a structured, interactive environment that renders hierarchical workflows as intuitive flow diagrams, linking donor acquisition, processing steps, and sequencing-ready outputs. As the resource provenance module of the NIMP digital twin, RPVE exposes the full dependency structure connecting physical tissue states to digital data products, thereby ensuring that the digital record accurately reflects laboratory processes. This direct alignment enables users to trace lineage, assess workflow integrity, and detect gaps or unexpected branching across centers.

RPVE supports targeted filtering, stepwise expansion of intermediate processes, and rapid inspection of resource-linked metadata, allowing researchers to verify provenance and assess cross-center consistency with minimal cognitive load. By maintaining an auditable and continuously synchronized representation of specimen-to-data transformations, RPVE resolves a central challenge in the digital twin paradigm: preserving structural clarity and factual accuracy as workflows scale. This capability enhances analytical rigor, facilitates collaborative oversight, and strengthens the reproducibility and transparency of the entire BICAN data production pipelines.

The RPVE user interface (Figure S6) consists of two main areas: the Sankey configuration panel (Figure S6.A) and the Sankey visualization panel (Figure S6.B). The configuration panel allows users to refine queries by selecting visualization direction (upstream or downstream), resource type, unique identifiers (NHash or local ID), maximum tracking depth, and label display. Based on these settings, the visualization panel dynamically generates a Sankey diagram representing provenance relationships, with nodes as resource items and edges as hierarchical connections. For example, Figure S6.B shows upstream relationships for a library pool (NHash ID LP-ROBCPL938294), tracing lineage from sequencing libraries back to donor tissues while revealing each processing stage, from dissection to library pooling. Interactive features, including mouse-hover highlights, click-based metadata access with JSON export, and zooming or download controls (Figure S6.B.1), enable multiple-level combination of interaction and exploration.

Role-based Access Control

NIMP employs a role-based access control (RBAC) framework to manage user permissions across both data and interface layers, ensuring secure and user-centric access to system functionalities. In this framework, permissions are defined by roles rather than by individual users, providing a scalable and transparent mechanism for implementing data governance and security policies. Each user in NIMP can be assigned one or multiple roles, and each role encapsulates a defined set of actions permitted within the system that range from data registration and curation to quality control and archive submission. To reflect the digital twin emphasis on intuitive, user-centered design, RBAC extends beyond data access management to include control over interface-level functionalities, ensuring that users encounter only the operations relevant to their roles and responsibilities.

Table 1 shows some of the important roles and their associated privileges in NIMP. For example, **Investigators** can search for brain slabs and submit specimen requests; **Brain Bank Personnel** can register donors and slabs, as well as review and approve requests; **Pre-approved Investigators** can directly enter externally authorized specimens; **Library Submitters** can register tissues and generate sequencing library packages; **Sequencing Center Personnel** can perform quality control and deposit FASTQ files; and **NeMO Personnel** can view completed sequencing outputs and contribute QC metrics for alignment-ready libraries. Collectively, this structured and auditable access framework ensures both operational efficiency and compliance with BICAN's multi-institutional data governance standards.

Role Name	**Privileges To Perform Actions in the NIMP**									
	Add Donor	Add Brain Slab	Review Request	Query Data	Request Specimen	Add Library	Upload Sequencing Result	Upload QC Metrics	Upload Aliquot Alignment QC	Upload Aliquot Archive URLs
Brain Bank Personnel	Y	Y	Y	Y						
Laboratory Investigator				Y	Y	Y				
Sequencing Core Personnel				Y			Y	Y		
NeMO Personnel				Y					Y	Y

Table 1. Key user roles and corresponding access privileges defined in the NIMP role-based access control framework

Unified Resource Browser

Unified Resource Browser (URB[23]) is an interactive data-access engine designed to enhance data exploration, FAIRness, transparency, and collaboration. As an integral component of the NIMP digital twin framework, URB provides a user-facing environment that mirrors and connects the digital representations of physical specimens and experimental workflows, allowing researchers to explore the evolving state of NIMP resources in real-time. It also functions as an integrative data exploration layer, bridging the Specimen Portal and the Sequence Library Portal to aggregate their data into a single interface. Through this unified interface, researchers can navigate eleven resource types and explore related model-card data without switching between portals or manually tracking identifiers through multiple data layers. For each resource, URB enables researchers to configure the corresponding model-card tables, perform targeted filtering and searches, export curated datasets, and leverage the integrated visualization module for deeper diver into digital assets. The browser builds on the RBAC to ensure data protection, granting users appropriate access to data based on their permissions while supporting collaborative activities. Advanced data table configuration management further facilitates users in creating, saving, and sharing personalized table views through direct links, fostering collaboration and reproducibility.

The Unified Resource Browser interface (Figure S7) is organized into three modules: user configuration; data table rendering; and data sharing, exporting, and visualization. The configuration module enables users to select a resource, apply project and lab filters, and choose relevant data columns for analysis. The data rendering module displays the configured dataset in an interactive table that updates dynamically as filters are adjusted. Users can further refine the dataset through keyword searches, sorting, numerical range filtering, and category-based selections, supporting targeted data exploration. These configurations can be saved, modified, or reloaded to support reproducible analyses. Configured data table can be shared via unique URL links that preserve all settings, exported as CSV files, or visualized directly within URB.

Unified Resource Browser covers eleven resource types aggregated from 17 projects in the BICAN community. As displayed in Table 2, the Donor resource contains the most extensive metadata, comprising 126 fields that describe 1,885 donors. The Library and Library Aliquot resources each contain over 20,000 records, reflecting the scale and richness of data accessible through NIMP. In addition to the eleven standard resource types, Unified Resource Browser enables researchers to create customized views that integrate data from additional resources. To build a customized view, users first select an anchor resource, specify the relevant data fields from that resource, and then incorporate joined columns from the remaining nine resources. The resulting cross-resource data table can be directly explored within the URB interface. This feature eliminates the need for time-consuming manual data aggregation and repeated identifier mapping offline, supporting exploration across the entire NIMP data hierarchy.

Table 2. Summary of resource items aggregated in Unified Resource Browser (October 2025)

Resource Type	Donor	Slab	Tissue	DCS[a]	ECS[b]	BCS[c]	AC[d]	Library	Library Aliquot	Library Pool
Metadata Fields	126	6	11	8	9	67	10	16	67	17
Total Count	1,885	9,702	19,387	8,239	8,928	16,980	13,876	21,917	22,085	1,107

[a]DCS: Dissociated Cell Sample; [b]ECS: Enriched Cell Sample; [c]BCS: Barcoded Cell Sample; [d]AC: Amplified cDNA.

Application Programming Interfaces

APIs have become a widely adopted technology for building web applications because they provide secure data communication between the system and external services, and function reusability for internal components. Built on a modern RESTful architecture, the NIMP APIs provides secure, flexible, and scalable data communication between the NIMP digital twin and external downstream services, supporting seamless access by data archives, computational pipelines, and the brain knowledge explorer. The APIs provide lookup of resource metadata instantly using NHash IDs or local identifiers, returning standard JSON-format documents that can be automatically consumed by external systems.

The NIMP API implements role-adapted data delivery that automatically adjusts information content based on user credentials. We use token-encoded role-based access control to enable real-time adaptation of responses to align with each user's workflow context, ensuring that brain bank personnel, laboratory investigators, sequencing centers, and archive managers each receive information structured for their specific activities while maintaining appropriate data privacy and integrity boundaries. The API also enables programmatic data uploading that streamlines larger-scale data ingestion by accepting structured experimental data in batch with built-in data validation checkpoints. It verifies whether the new data matches the uploader's role, evaluates data completeness against resource model-cards, checks for duplicate entries, confirms anatomical and ontological term accuracy, and commits digital resource creation according to pre-defined relationships. These capabilities ensure that NIMP's digital representation remains continuously synchronized with ongoing laboratory activities, creating a unified view of BICAN resource states through the digital twin paradigm across the entire community.

Change-log Dashboard, Data Auditing and Quality Assurance

NIMP provides an integrated change-log dashboard that centralizes all audit and provenance information to support continuous data quality assurance across the tissue-to-bytes workflow. Every action within the system, from creating tissue records to editing cell-sample metadata or updating sequencing-library entries, is automatically logged with timestamps, user identifiers, and contextual metadata. This persistent, queryable dashboard allows laboratory teams to monitor workflow activities, identify discrepancies, and validate that protocol-specific steps were executed correctly. At the consortium level, the NeMO archive further aggregates and cross-checks these logs, ensuring that all modifications propagated through NIMP retain their provenance integrity. By embedding fine-grained change tracking and automated quality-control measures in their corresponding locations within the digital twin, NIMP embeds provenance auditing into a discipline of operational quality assurance, reinforcing FAIRness principles and preserving the trustworthiness of all resource types.

REFERENCES

1. Brain Initiative Cell Atlas Network https://braininitiative.nih.gov/research/tools-and-technologies-brain-cells-and-circuits/brain-initiative-cell-atlas-network
2. Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Initiative - Cell Census Network (BICCN) https://biccn.org/
3. BICAN R21 RFA https://grants.nih.gov/grants/guide/rfa-files/RFA-MH-21-237.html
4. BICAN UM01RFA https://grants.nih.gov/grants/guide/rfa-files/RFA-MH-21-235.html
5. https://www.portal.brain-bican.org/

6. Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016).
7. F. Tao and M. Zhang, Digital Twin Shop-Floor: A New Shop-Floor Paradigm Toward Smart Manufacturing, IEEE Access, vol. 5, pp. 20418–20427, 2017.
8. M. Viceconti and P. Hunter, The Virtual Physiological Human: Ten Years After, Interface Focus, vol. 6, no. 2, 2016.
9. K. Amunts et al., The Human Brain Project: A European Research Infrastructure for Brain Research, Neuron, vol. 110, no. 13, pp. 2234–2259, 2022.
10. EBRAINS Infrastructure Overview, EBRAINS, https://www.ebrains.eu/
11. S. Van Sint Jan, L. Geris, Modelling towards a more holistic medicine: The Virtual Physiological Human (VPH), Morphologie, Volume 103, Issue 343, 2019, Pages 127-130, ISSN 1286-0115, https://doi.org/10.1016/j.morpho.2019.10.044.
12. NIH Common Fund, Bridge2AI Program Overview, accessed 2025. https://bridge2ai.org/
13. Health Data Research UK, Trusted Research Environments and National Data Infrastructure, HDR UK, accessed 2025.
14. https://datareadiness.eng.ox.ac.uk/projects/ (IMI)
15. https://www.cde.ual.es/documentacion/bases-de-datos/ (EOSC-Life)
16. Stober, T., Hansmann, U. (2010). Traditional Software Development. In: Agile Software Development. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-70832-2_2
17. Hirsch JS, Tanenbaum JS, Lipsky Gorman S, Liu C, Schmitz E, Hashorva D, et al. Harvest, a longitudinal patient record summarizer. J Am Med Inform Assoc. 2015;22(2):263–74.
18. Zhang GQ, Tao S, Xing G, Mozes J, Zonjy B, Lhatoo S. D, Cui L, et al. NHash: randomized n-gram hashing for distributed generation of validatable unique study identifiers in multicenter research. JMIR medical informatics. 2015; 3(4): e4959.
19. Ruby on Rails [Internet]. Ruby on Rails. Available from: https://rubyonrails.org/. Access Date: 25 March, 2025.
20. Highcharts [Internet]. Available from: https://www.highcharts.com/. Access Date: 25 March, 2025.
21. Michael Hawrylycz et al., A Community Standard Multispecies Cell Atlas of the Basal Ganglia: The BRAIN Initiative Cell Atlas Network, submitted in the same package.
22. Li X, Huang Y, Ng L, Smith K, Chou WC, Abeysinghe R, Tong L, Lin S, Cui L, Tao S, Zhang GQ. Relational Database-Based Resource-Provenance Visualization Engine: with an application to BICAN data. In AMIA Annual Symposium Proceedings 2025. American Medical Informatics Association.
23. Lin S, Tong L, Smith K, Ng L, Mollenkopf T, Huang Y, Abeysinghe R, Chou WC, Byrd A, Cui L, Zhang GQ, Li X, Tao S. Unified Resource Browser: An Interactive and Reconfigurable Web Framework for Biomedical Metadata Exploration, Visualization, and Management. In AMIA Annual Symposium Proceedings 2025. American Medical Informatics Association.
24. Chen KH, Boettiger AN, Moffitt JR, Wang S, Zhuang X. Spatially resolved, highly multiplexed RNA profiling in single cells. Science. 2015 Apr 24;348(6233):aaa6090.