Human–AI Collaboration for Research Data Management in Construction

Leveraging OpenConstruction and Human-Centered Data Work to Advance FAIR (Findable, Accessible, Interoperable, Reusable) and Open Science

Pingbo Tang, Ph.D. P.E., Associate Professor

Human-Machine Harmony for Infrastructure Laboratory

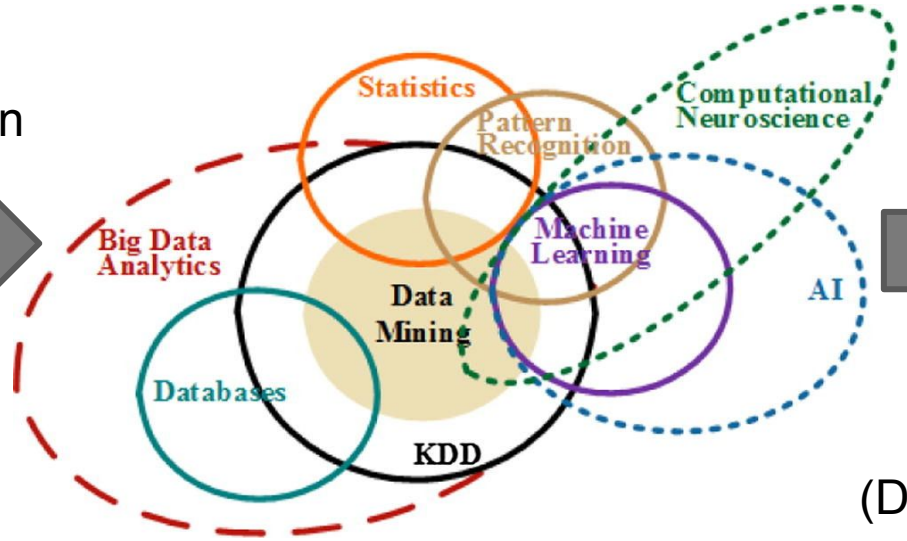
NSF Cyberinfrastructure for Public Access and Open Science (CI PAOS) Workshop on Research Data Management in Construction | May 7–8, 2026 | Michigan Technological University

Talk Outline

- **Motivation:** The imperative and complexities of research data management in the construction sector
- **Evidence:** Insights from a practitioner survey on perceived barriers and promotion strategies
- **Platform:** OpenConstruction and the FAIR principles as a foundational reference ecosystem for Research Data Management in Construction (RDMC)
- **Human Data Work:** Leveraging AI agents to observe, learn from, and enhance human modeling and sharing practices throughout the facility lifecycle
- **Research Agenda:** Posing research questions and identifying collaborative opportunities within the scope of NSF CI PAOS

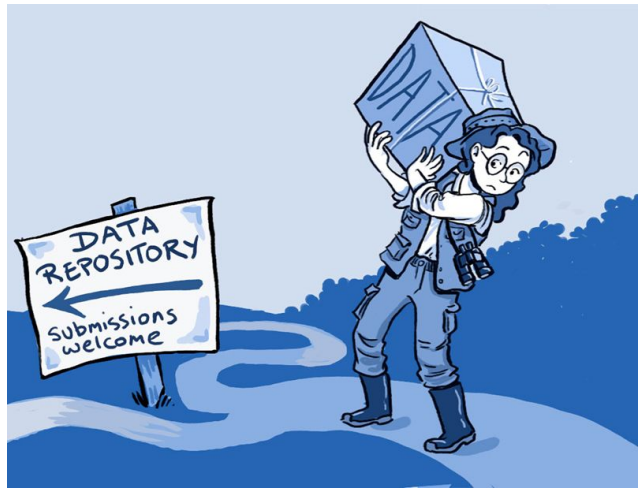
Increasing Data in Civil and Construction Engineering

Increasing Generated Data in Civil Engineering
- BIM
- AR/VR
- Monitoring Data
- …



Better Civil Engineering
- Clash detection and resolution
- Performance prediction
- Facility management
- …

(Dean et al., 2014)



- Construction projects now generate BIM models, AR/VR simulations, UAV surveys, IoT streams, digital twins, and unstructured documents — all in fragmented silos.
- Better data sharing could transform clash detection, predictive maintenance, and facility management — but barriers remain.
  - Privacy, Technology, Cost-Benefit, …

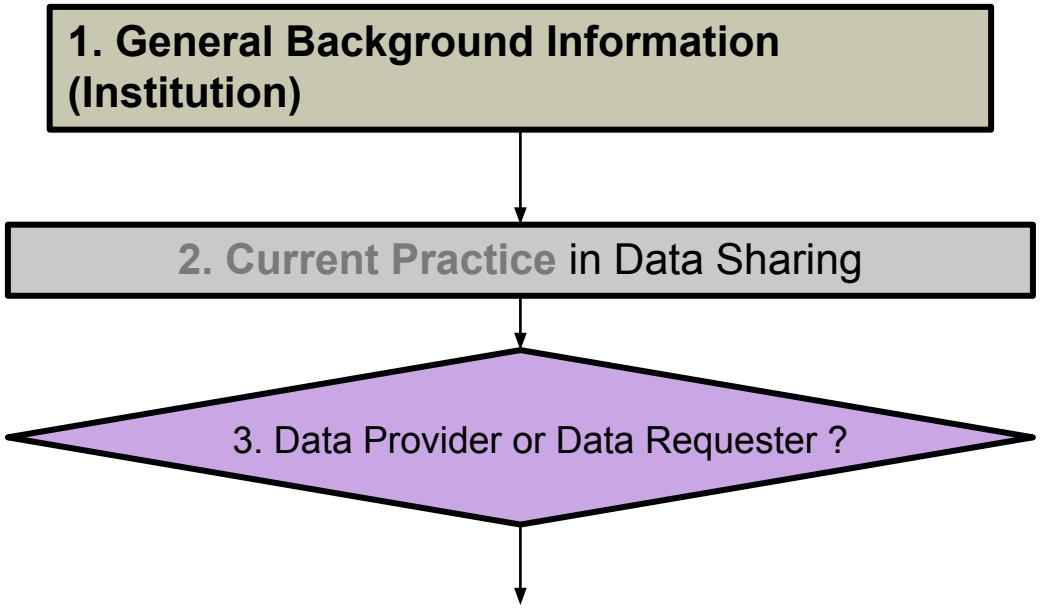


What Does Research Data Management in Construction Actually Cover?

- **Data types:** BIM and CAD models; images and videos from cameras, UAVs, and wearables; IoT sensor streams; simulation and digital twin outputs; inspection reports and maintenance logs
- **Lifecycle scope:** Spans design → construction → operation → maintenance of facilities — data created in each phase must remain interpretable and reusable in the next
- **Beyond datasets:** RDMC also includes user provenance — records of how designers, engineers, and field workers model, analyze, annotate, and hand off data inside their software tools
- **The goal:** Make all of this Findable, Accessible, Interoperable, and Reusable (FAIR) across organizations and over the full facility lifecycle

The work presented in the following slides was completed by the ASCE Computing Division. However, these materials do not reflect the official opinion of ASCE, and the speaker assumes full responsibility for the thoughts expressed herein.

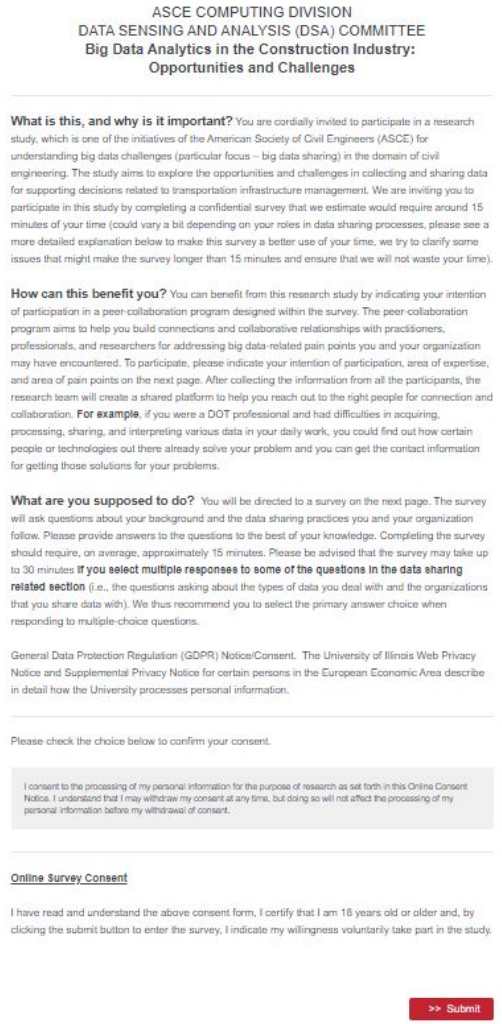
Survey of Data Sharing Practices in Civil Infrastructure Engineering



- 4. Challenge in Data Sharing
- 5. Future Vision in Data Sharing
- 6. Promotion Strategies Preference in Data Sharing

ASCE
AMERICAN SOCIETY OF CIVIL ENGINEERS

- "Qualtrics survey conducted February–August 2020 with researchers, practitioners, and agency staff"
- "Covered: current practice, perceived barriers, future vision, and promotion strategy preferences"
- "Published: Wang, Y., Tang, P., et al. (2023). Journal of Computing in Civil Engineering, 37(2)"



ASCE COMPUTING DIVISION
DATA SENSING AND ANALYSIS (DSA) COMMITTEE
Big Data Analytics in the Construction Industry: Opportunities and Challenges

What is this, and why is it important? You are cordially invited to participate in a research study, which is one of the initiatives of the American Society of Civil Engineers (ASCE) for understanding big data challenges (particular focus – big data sharing) in the domain of civil engineering. The study aims to explore the opportunities and challenges in collecting and sharing data for supporting decisions related to transportation infrastructure management. We are inviting you to participate in this study by completing a confidential survey that we estimate would require around 15 minutes of your time (could vary a bit depending on your roles in data sharing processes, please see a more detailed explanation below to make this survey a better use of your time, we try to clarify some issues that might make the survey longer than 15 minutes and ensure that we will not waste your time).

How can this benefit you? You can benefit from this research study by indicating your intention of participation in a peer-collaboration program designed within the survey. The peer-collaboration program aims to help you build connections and collaborative relationships with practitioners, professionals, and researchers for addressing big data-related pain points you and your organization may have encountered. To participate, please indicate your intention of participation, area of expertise, and area of pain points on the next page. After collecting the information from all the participants, the research team will create a shared platform to help you reach out to the right people for connection and collaboration. **For example**, if you were a DOT professional and had difficulties in acquiring, processing, sharing, and interpreting various data in your daily work, you could find out how certain people or technologies out there already solve your problem and you can get the contact information for getting those solutions for your problems.

What are you supposed to do? You will be directed to a survey on the next page. The survey will ask questions about your background and the data sharing practices you and your organization follow. Please provide answers to the questions to the best of your knowledge. Completing the survey should require, on average, approximately 15 minutes. Please be advised that the survey may take up to 30 minutes **if you select multiple responses to some of the questions in the data sharing related section** (i.e., the questions asking about the types of data you deal with and the organizations that you share data with). We thus recommend you to select the primary answer choice when responding to multiple-choice questions.

General Data Protection Regulation (GDPR) Notice/Consent. The University of Illinois Web Privacy Notice and Supplemental Privacy Notice for certain persons in the European Economic Area describe in detail how the University processes personal information.

Please check the choice below to confirm your consent.

I consent to the processing of my personal information for the purpose of research as set forth in this Online Consent Notice. I understand that I may withdraw my consent at any time, but doing so will not affect the processing of my personal information before my withdrawal of consent.

Online Survey Consent

I have read and understand the above consent form, I certify that I am 18 years old or older and, by clicking the submit button to enter the survey, I indicate my willingness voluntarily take part in the study.

>> Submit

Different Numbers of Responses in Different Sections of Survey

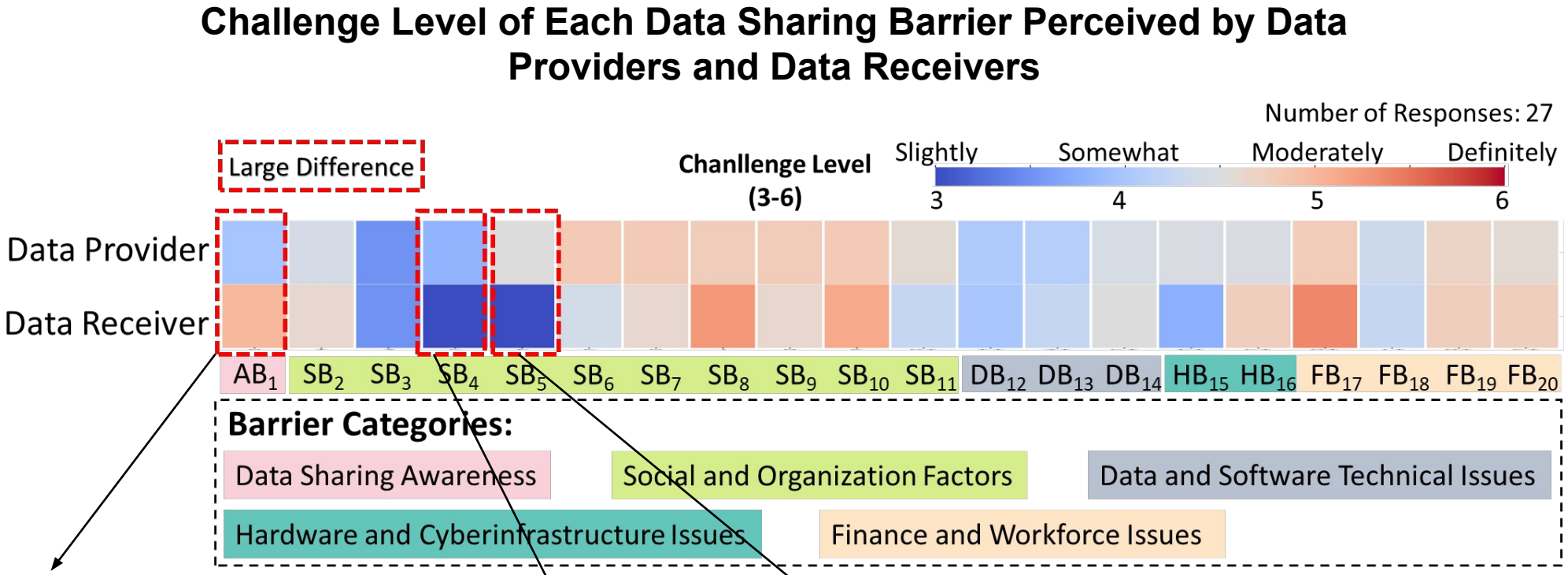
Section Name	Current Practice	Future Vision	Barriers	Data-sharing Promotion Strategies
Response Count	45	56	20	25

Perceived Barriers to Data Sharing in Civil Infrastructure

Data Sharing Barriers in the Civil Infrastructure Engineering

Data Sharing Awareness (AB):
(1) There is a lack of awareness of existing platforms/avenues for sharing or requesting data.
Social and Organization Factors (SB):
(2) Others would use my data to publish before me (3) Others would criticize my work
(4) Others would falsify my work (5) The data could be misinterpreted
(6) Data are sensitive (7) Data are private
(8) Data are of competitive edges
(9) Lack of standard, policy, or consensus regarding data sharing
(10) Ethical and legal issues causing concerns about data ownership, legal implications, and intellectual property (both for individuals and organizations)
(11) Lack of comprehensive data security technology support
Data and Software Technical Issues (DB):
(12) The data structure is too complex to the extent of making data sharing worthless (e.g., others do not know how to use the data)
(13) Lack of quantification of data quality
(14) Lack of automatic technology for extraction, integration, and analysis of information from various data formats, such as integrated analysis of images and texts
Hardware and Cyberinfrastructure Issues (HB):
(15) Cyberinfrastructure could not support big data use (e.g., limited internet bandwidth, storage space, computing power, and so on) (16) data access is not available
Finance and Workforce Issues (FB):
(17) Major efforts are required to collect the data (18) Major efforts are required to share the data
(19) Engineering/technical support for data sharing is limited
(20) Financial and human resources with a computational mindset for data sharing are limited

Perceived Barriers to Data Sharing in Civil Infrastructure

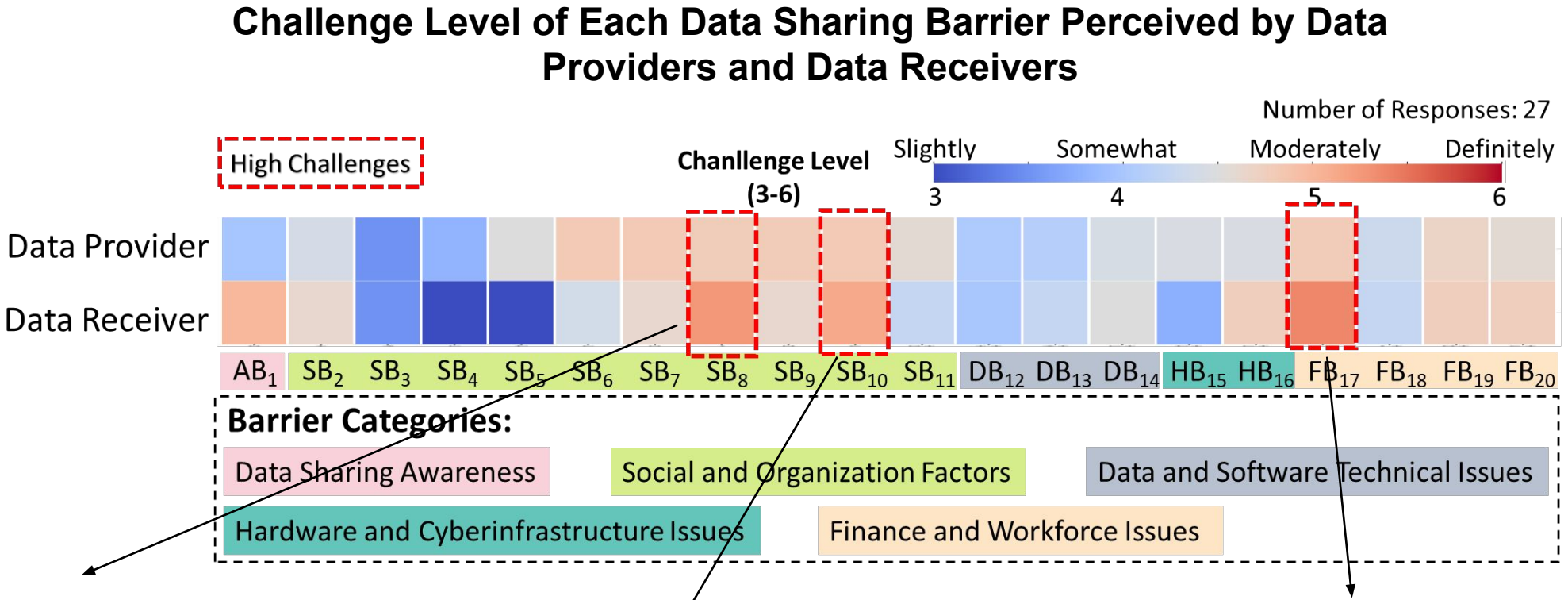


There is a lack of awareness of existing platforms/avenues for sharing or requesting data.

Others would falsify my work

The data could be misinterpreted

Perceived Barriers to Data Sharing in Civil Infrastructure



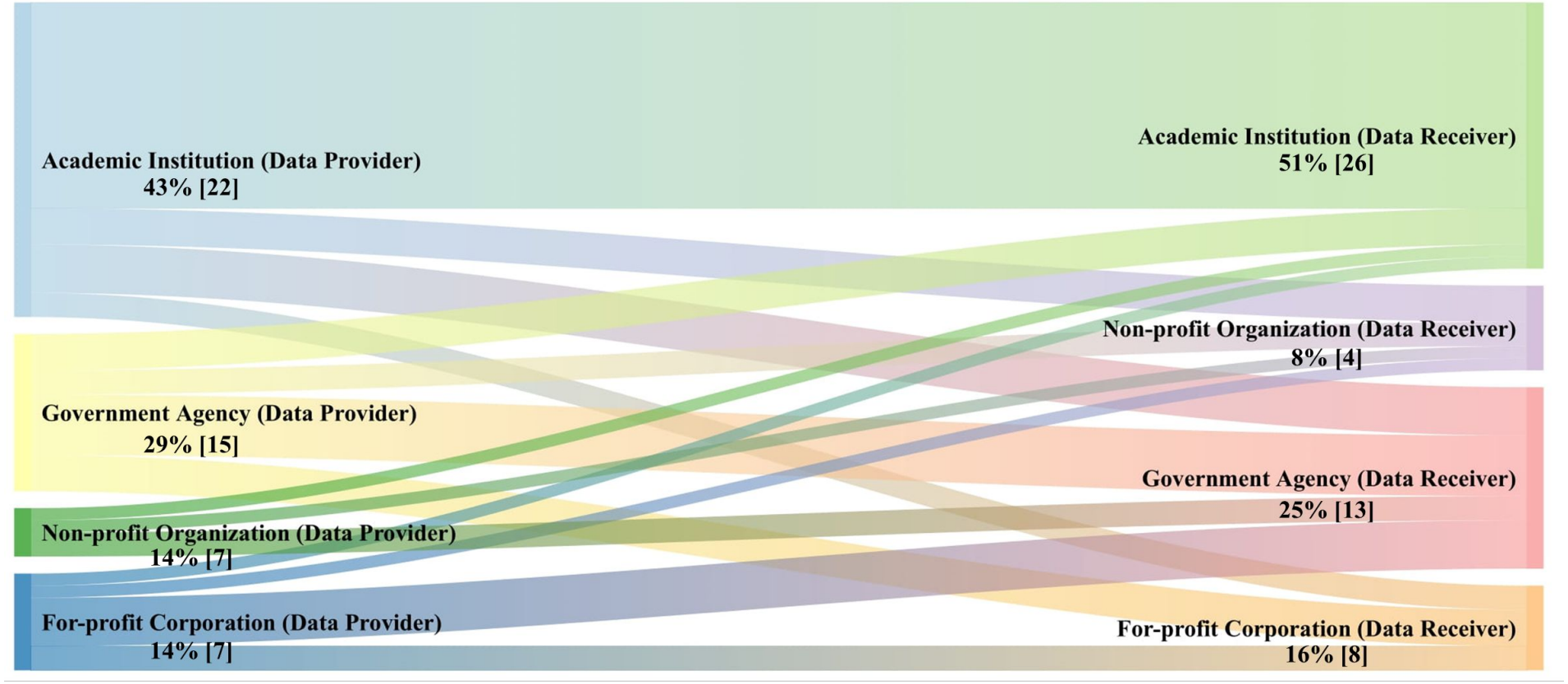
Data are of competitive edges

Major efforts are required to collect the data

Ethical and legal issues causing concerns about data ownership, legal implications, and intellectual property (both for individuals and organizations)

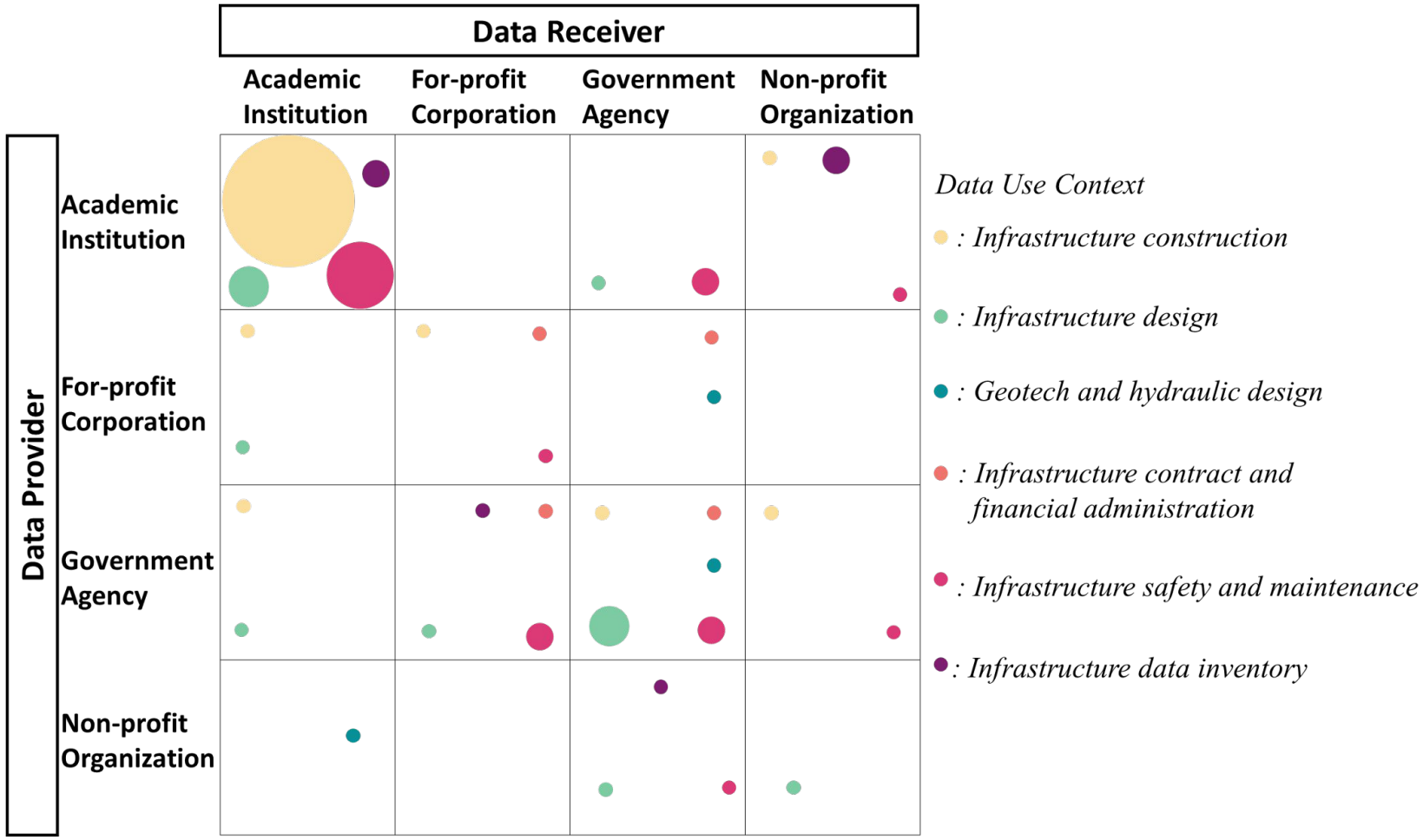
Current Data Sharing Practice Varies Widely Across Organizations

The Strength of Current Data Sharing Practice among Different Organizations

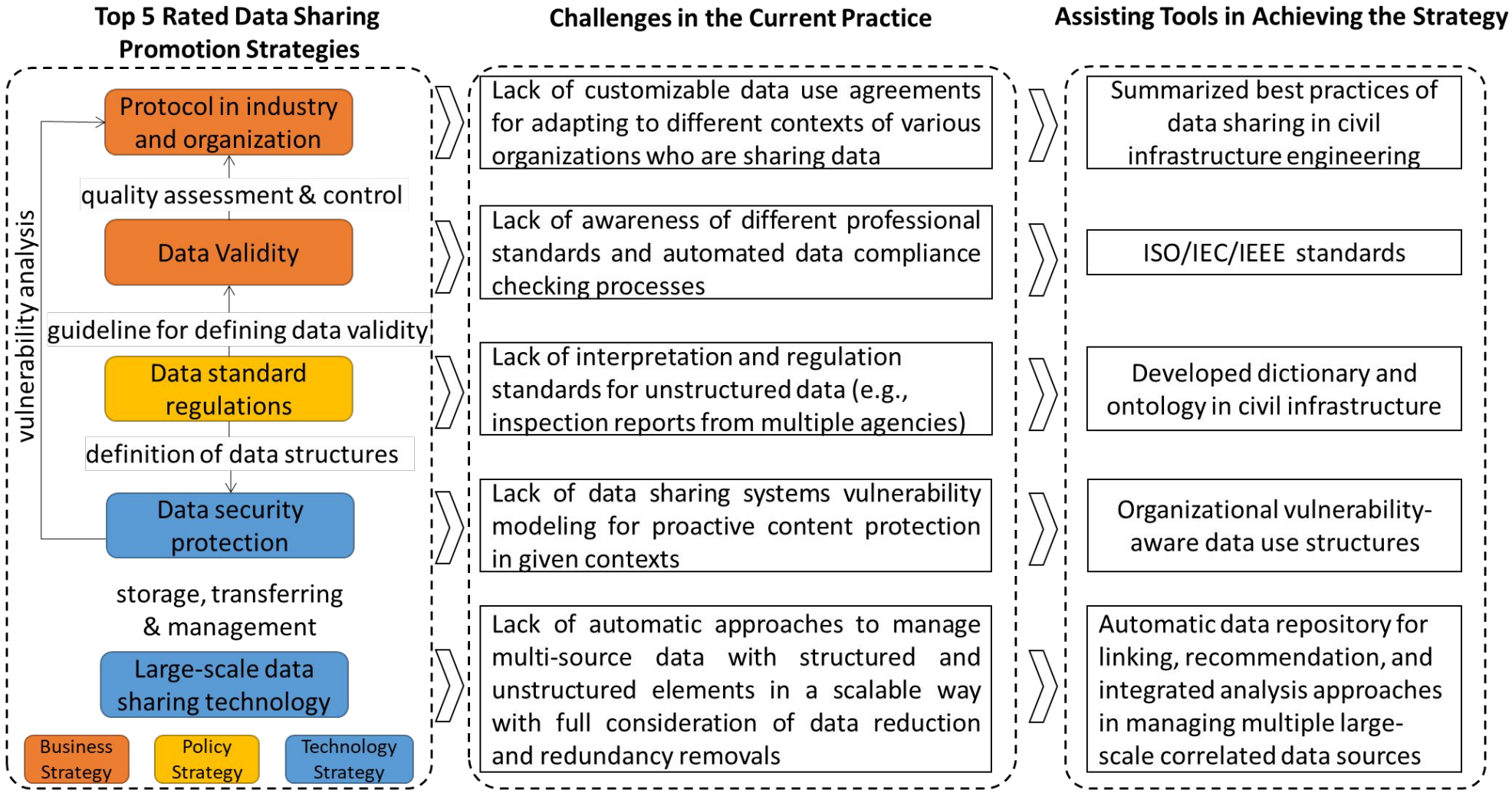


Current Data Sharing Practice Varies Widely Across Organizations

Number of Data Use Contexts in Responses from Different Types of Organizations Participated in Data Sharing



Technology, Policy, Business, and Academic Roadmaps for Data Sharing



Technology, Policy, Business, and Academic Roadmaps for Data Sharing

- **Technology/Academic:** Platforms, metadata standards, cybersecurity, and data validation protocols
- **Policy/Government:** Data ownership frameworks, legal protections, IP and licensing guidelines
- **Business/Industry:** Incentives for sharing, value creation from reuse, confidentiality safeguards
- **Cross-cutting:** Training and cultural change for researchers, engineers, and field workers

ASCE
AMERICAN SOCIETY OF CIVIL ENGINEERS

From Data Sharing Challenges to RDMC and AI

- **The survey shows that fragmentation, mistrust, and effort are the main barriers**
  - not just lack of technology
- **Solving RDMC requires infrastructure that supports FAIR principles and rich provenance, not just better file storage or transfer**
- **RDMC must manage two things together:**
  - (1) the datasets themselves, and (2) user provenance — how humans model, interpret, and exchange data inside their tools across the facility lifecycle
- **AI agents that observe these human data work patterns can help synthesize good practices, flag risks, and propose new sharing principles that respect heterogeneous domain terms**
- **Next: OpenConstruction as a reference design for this kind of RDMC ecosystem**

FAIR: Findable, Accessible, Interoperable, Reusable

- **Findable:** Data and workflows have persistent identifiers and rich, searchable metadata so others can reliably locate them.
- **Accessible:** Clear, documented access paths and permissions for both humans and machines (e.g., portals, APIs, license terms).
- **Interoperable:** Shared schemas, ontologies, and formats allow combining datasets and tools across projects, organizations, and domains.
- **Reusable:** Data, models, and workflows include sufficient documentation, provenance, and licensing to support new questions beyond the original study.

What FAIR Looks Like in Construction Research

- **Findable:** BIM and point cloud datasets with persistent identifiers, documented coordinate systems, version history, and links to related inspection or sensor data.
- **Accessible:** Construction image datasets such as the Advanced Construction Image Dataset Suite (ACID), which provide tens of thousands of annotated images of construction machines with clear download procedures, licenses, and documented tasks (e.g., object detection, segmentation, captioning).
- **Interoperable:** Digital twins that link models, IoT streams, and simulation outputs through shared identifiers, schemas, and units so they can be combined across tools and organizations.
- **Reusable:** Published workflows and scripts with recorded parameters and provenance so other teams can rerun analyses on new projects or extended datasets.

Domain-Specific RDMC Principles for Construction

Domain-Specific Principles

- **Context-aware**: what assets, activities, and conditions were captured?
- **Risk-aware**: privacy, safety, security implications of sharing data (e.g., worker images)
- **Task-aware**: what downstream tasks are supported?

FAIR Data Principles

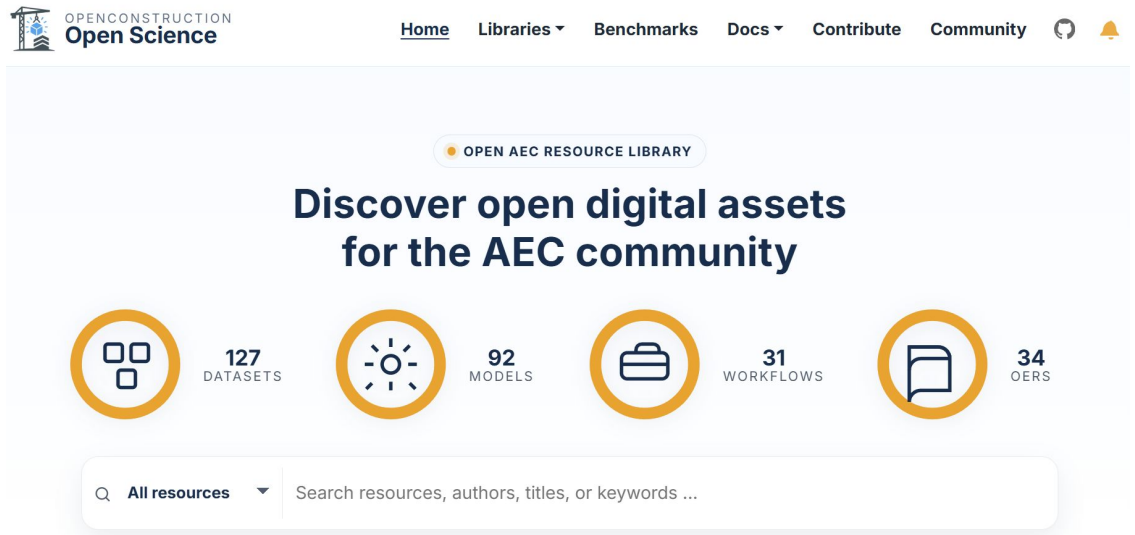
- **F – Findable**
  - Unique identifiers, searchable metadata
- **A – Accessible**
  - Clear access paths
- **I – Interoperable**
  - Common schemas, ontologies, formats
- **R – Reusable**
  - Rich descriptions, clear licenses

OpenConstruction applies FAIR plus these construction-specific principles to real AEC datasets, including ACID and other visual benchmarks (Xiong et al., 2026).

OpenConstruction Initiative: A Socio-Technical Partnership for RDMC and Open Science in AEC

- **VISION:** A community-driven open-science platform that indexes digital assets for the AEC community, connecting AEC domain researchers, educators, and community stakeholders
- **GOALS:** Develop and validate methods that improve how AEC digital assets are accessed, discovered, and reused
  - Bridge research, industry, and education through shared resources
  - Provide a collaborative infrastructure for future AEC foundation models and AI tools

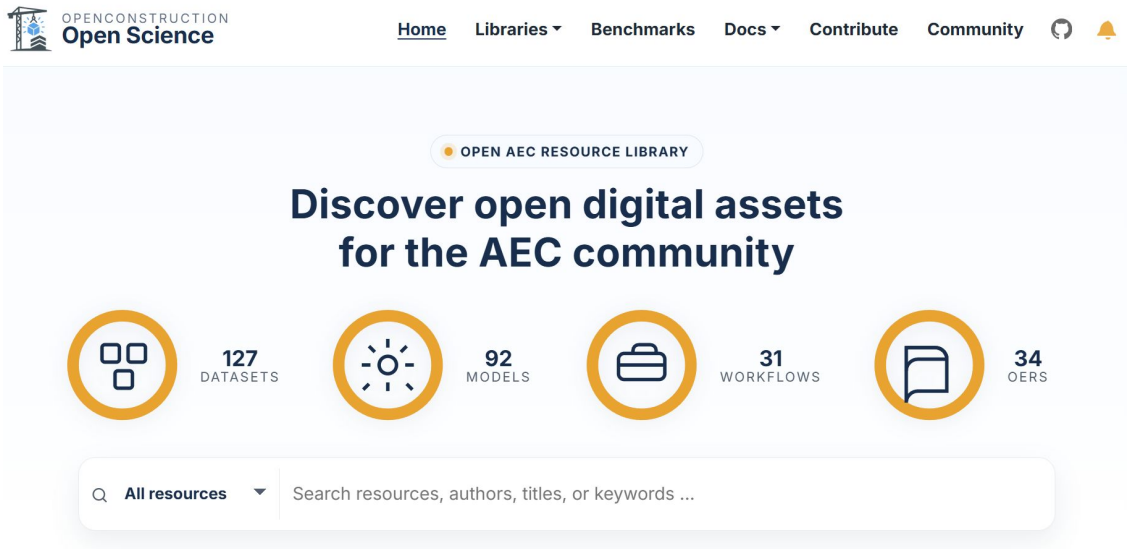




(Xiong et al., 2026) OpenConstruction Open Science

Real-World Platforms: FAIR + Construction-Aware RDMC Principles





OpenConstruction applies FAIR plus these construction-specific principles to real AEC datasets, including ACID and other visual benchmarks.

Domain-Specific Principles

- **Context-aware**: what assets, activities, and conditions were captured?
- **Risk-aware**: privacy, safety, security implications of sharing data (e.g., worker images)
- **Task-aware**: what downstream tasks are supported?

FAIR Data Principles

- **F – Findable**
  - Unique identifiers, searchable metadata
- **A – Accessible**
  - Clear access paths
- **I – Interoperable**
  - Common schemas, ontologies, formats
- **R – Reusable**
  - Rich descriptions, clear licenses

Metadata Model for Context-Rich Construction Datasets

Metadata Element	Proposed	Dublin Core	Google Dataset	Zenodo	Roboflow	Hugging Face
General Metadata						
Identifier / URL	✓	✓	✓	✓	✓	✓
Title	✓	✓	✓	✓	✓	✓
Authors / Contributors	✓	✓	✓	✓	✓	✓
Release Date	✓	✓	✓	✓	✓	✓
License	✓	✓	✓	✓	✓	✓
DOI (if available)	✓	–	✓	✓	✓	✓
Source	✓	✓	✓	✓	✓	✓
Scholarly Records	✓	–	✓	✓	✓	✓
Data / Model Preview	✓	–	–	✓	✓	✓
Dataset-Specific Metadata						
Dataset Size	✓	✓	○	✓	✓	✓
Data Modality	✓	✓	○	○	✓	✓
Annotated Classes	✓	○	○	○	✓	○
Annotation Types	✓	○	○	○	✓	○
Resolution	✓	○	○	○	○	○
Domain Tasks	✓	○	○	○	✓	✓
Collection Location	✓	○	✓	○	○	○
Dataset Link	✓	✓	✓	✓	✓	○
Model-Specific Metadata						
Input Modality	✓	✓	–	○	○	○
Domain Tasks	✓	○	–	○	✓	✓
Application Domain	✓	○	–	○	○	○
Description (abstract)	✓	✓	–	✓	✓	✓
Model Link	✓	✓	–	✓	✓	✓

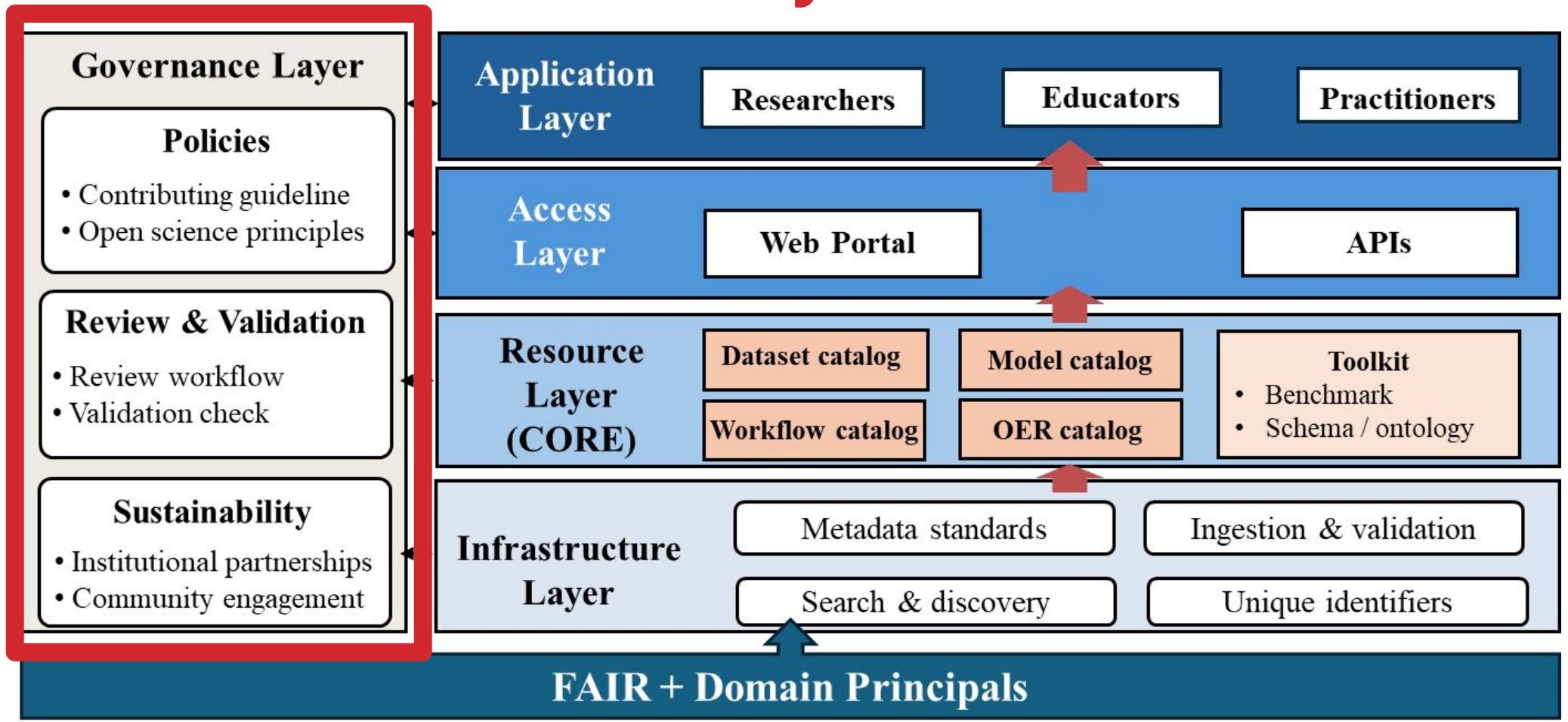
Encodes scene, activity, sensor, and task information that downstream users and AI models need to interpret and reuse datasets (e.g., ACID/ACIDB scenes, equipment classes, camera setups).

Highlights gaps where information is only in unstructured descriptions — targets for future RDMC improvements (e.g., standardizing safety/privacy annotations or task definitions).

- ✓: structured and queryable;
- O: partially supported or typically embedded in unstructured text (e.g., descriptions, tags);
- –: not explicitly supported

(Xiong et al., 2026)

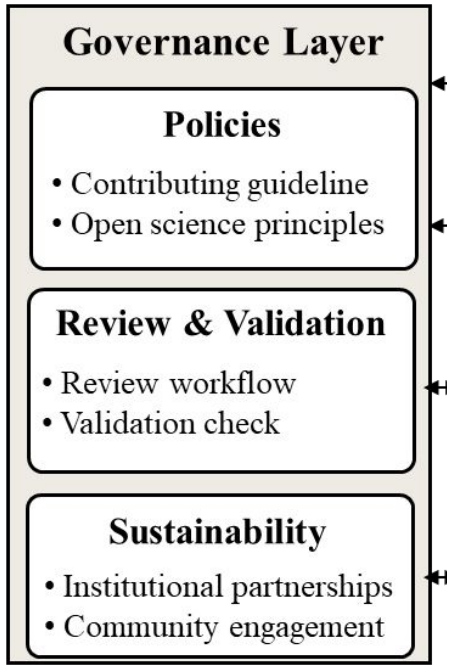
Layered Architecture for a Construction RDMC Ecosystem



Ingest → Curate → Catalog → Access → Reuse → Feedback

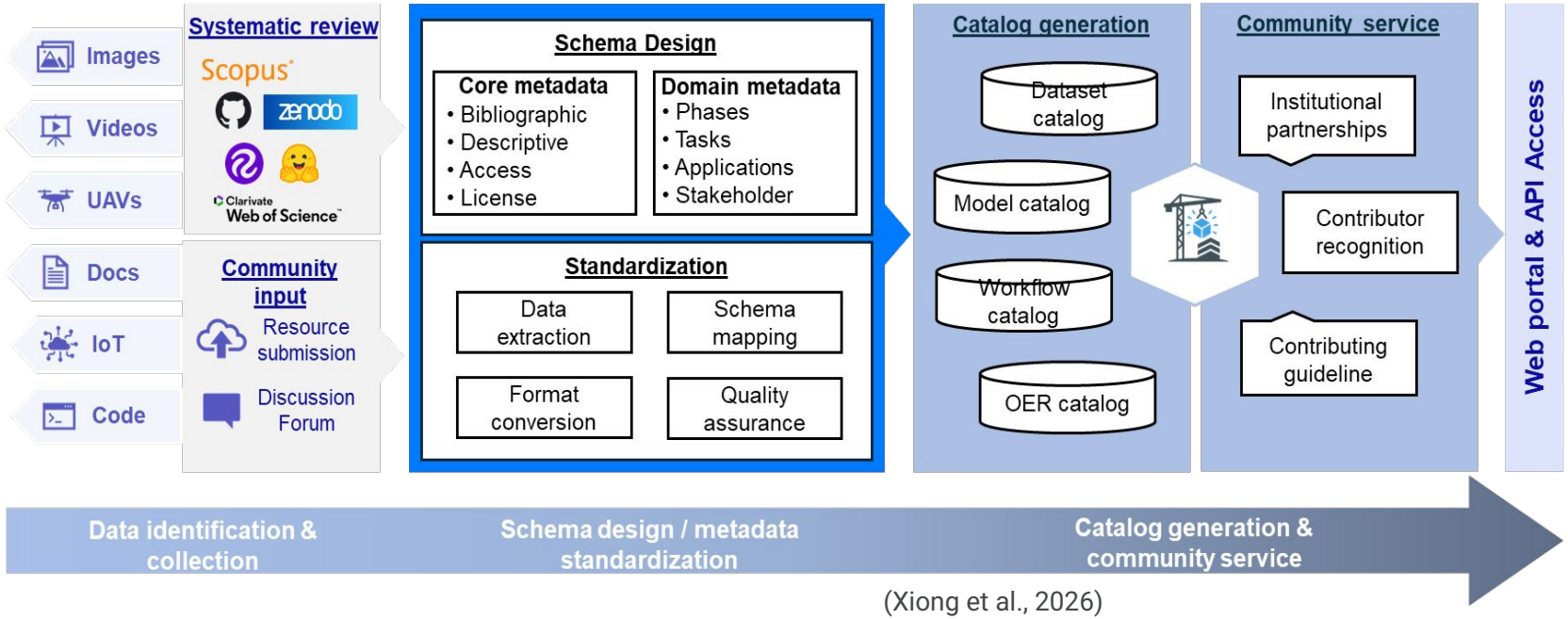
(Xiong et al., 2026)

Governance: Policies, Validation, and Sustainability



- **Policies:** contribution guidelines, open science principles, privacy/security rules for visual datasets like ACID.
- **Review & Validation:** review workflows and dataset validation checks (technical quality, documentation completeness).
- **Sustainability:** institutional partnerships (e.g., with dataset owners and CI PAOS projects), community engagement.

Governing the process of integrating resources into OpenConstruction



(Xiong et al., 2026)

From Open Catalogs to Certified Knowledge for AI

- **OpenConstruction and datasets like ACID show that we can build FAIR, domain-aware catalogs of AEC data, workflows, and OER.**
- **However, AI agents need more than ‘open’:**
  - “They” need structured, certified units of knowledge with clear provenance, validity, and usage rules.
- **The Tokenization for Engineering Knowledge initiative:**
  - Treats each standard, dataset, workflow—or even human data work trace—as a tokenized unit that can be verified, attributed, and safely consumed by AI co-pilots.

Ongoing work that may not reflect the official opinion of ASCE and the speaker takes the responsibility of these expressed thoughts.

ASCE
TECHNICAL ADVANCEMENT AI TASK FORCE

Tokenization for Engineering Knowledge: Overview

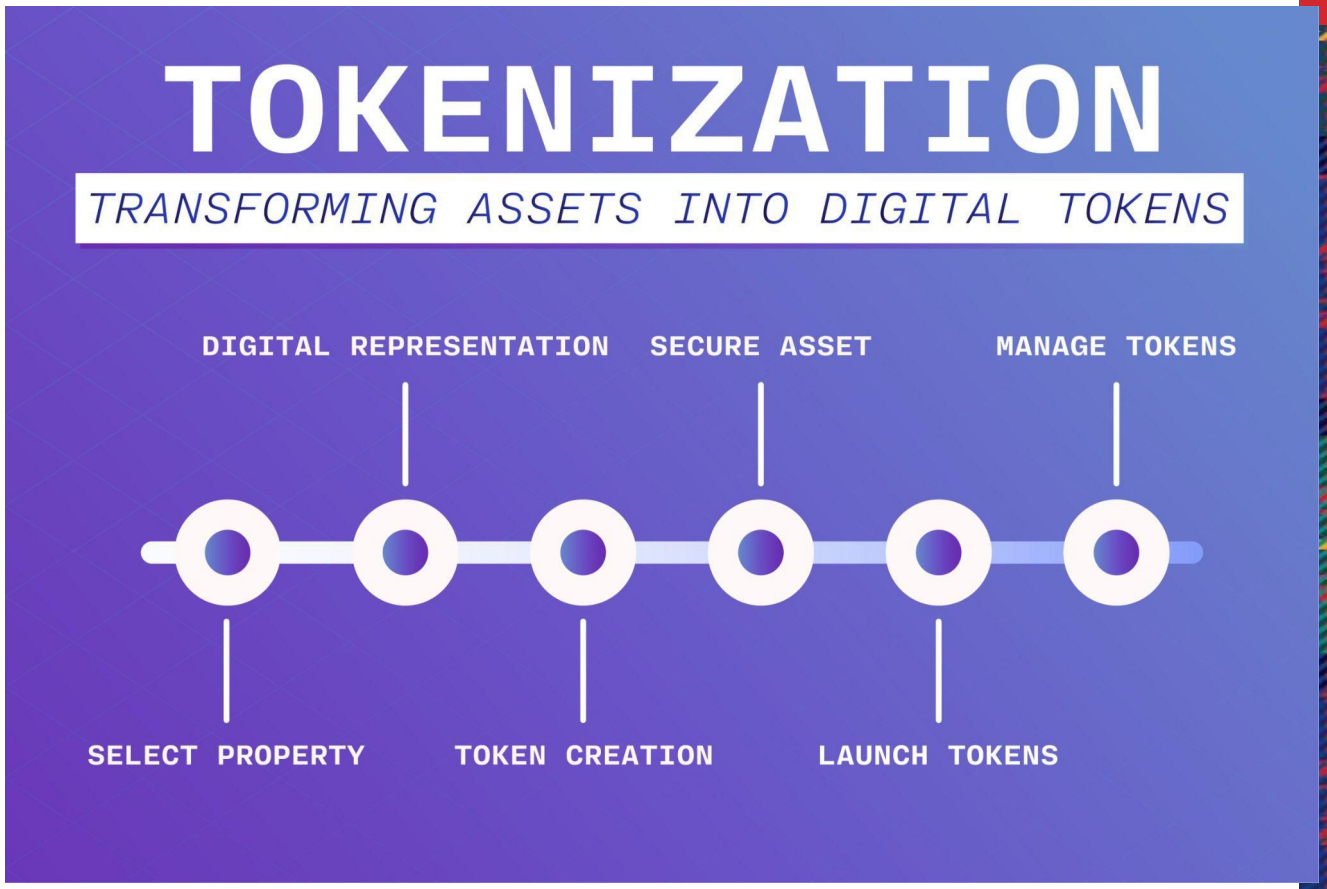
ASCE TECHNICAL ADVANCEMENT AI TASK FORCE

- The **Tokenization for Engineering Knowledge** initiative aims to help ASCE evolve from a publisher of static documents into **a trusted steward of civil engineering knowledge in the AI and digital economy**.
- The committee's goal is to design **a decentralized yet governable ecosystem for engineering data and knowledge reuse**, where provenance, attribution, verification, and responsible AI integration are built in from the start.



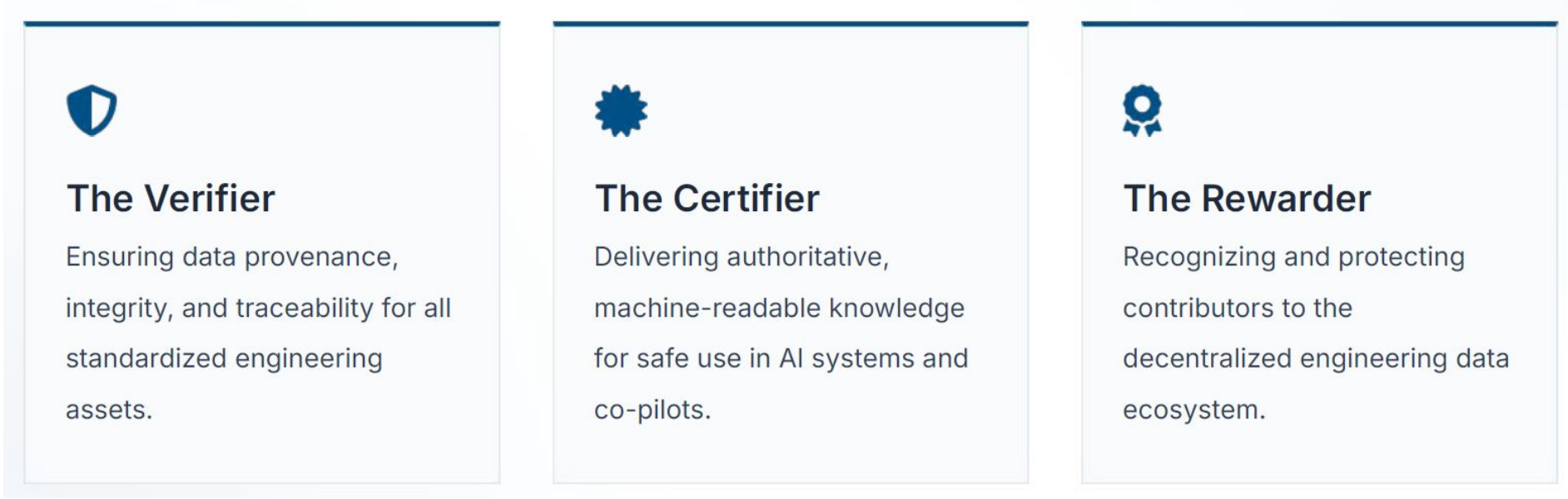
What Is Tokenization?

- **Converting standards, guidelines, and domain models into secure, machine-readable knowledge packages.**
- **Structured Units:**
  - Breaking down complex standards into specific, usable data points.
- **Standardized Metadata:**
  - Creating clear definitions, domains, and relationships.
- **Cryptographic Provenance:**
  - Utilizing hash-tree structures (e.g., Merkle trees) to guarantee integrity and detect tampering.
- **AI Integration:**
  - Allowing AI models to rely on ASCE-certified tokens as absolute sources of truth, reducing hallucination.



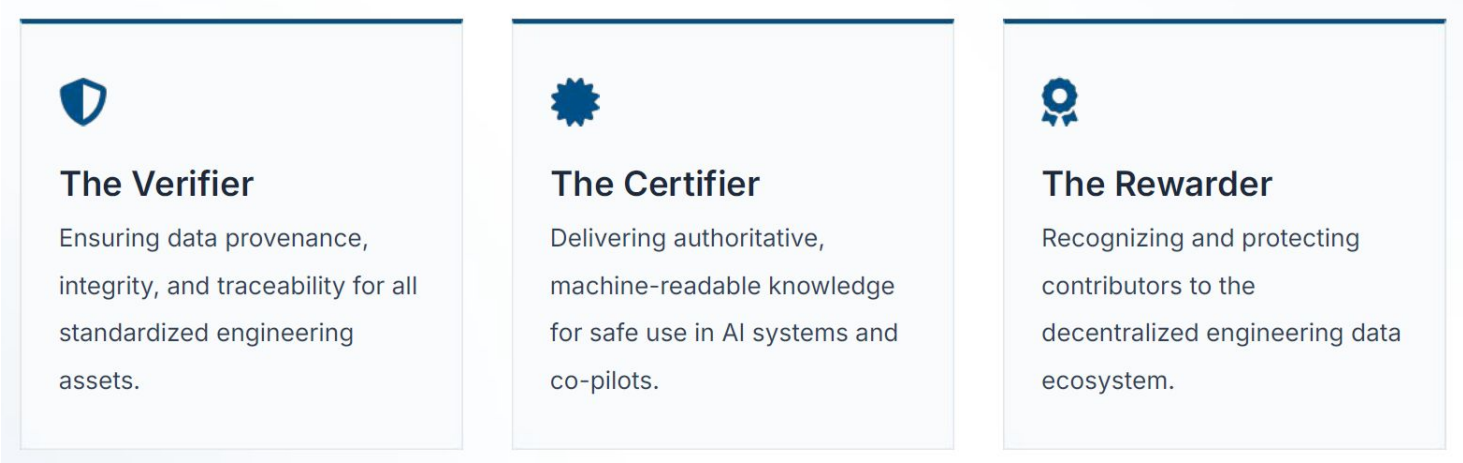
ASCE Cyberinfrastructure for Sustainable Growth of AEC Knowledge Economy?

ASCE should evolve from being a publisher of static documents to being the trusted steward of civil engineering knowledge in the AI / digital economy, coordinating various data and knowledge system users by tracing their works and giving them tokens for their works in helping the growing of AEC data and knowledge ecosystem.



Sustainable Decentralized AEC Data and Knowledge Business Model?

ASCE TECHNICAL ADVANCEMENT AI TASK FORCE



- Everyone can get tokens by using and contributing to the data and knowledge ecosystem: contributing data, workflows, verifying them, and providing education resources.
- More successful contributions and verifications produce more tokens, encouraging everyone to make the data and knowledge ecosystem more reliable and widely available.
- A carefully designed cybersecurity and governance model is key for the healthy growth of such an ecosystem.
- **ASCE would ideally coordinate the cyberinfrastructure and collect a "tax," but the individual tokens belong to the real contributors of data, knowledge, models, workflows, and educational resources.**

Call for Participation: ASCE Tokenization Ecosystem Workshop & Hackathon Showcase

A Proposed Hackathon Initiative to be announced in the coming months.

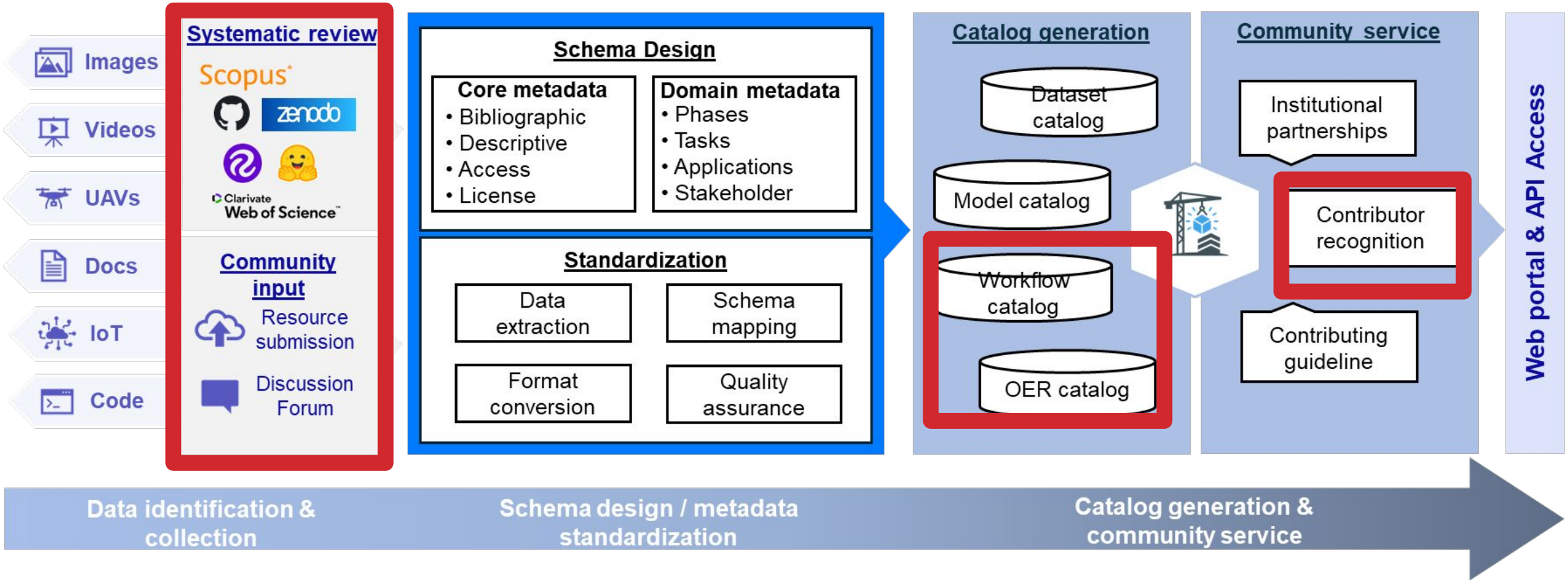
- Design the ASCE Data and Knowledge Tokenization Solution.
  - How you would like to transform civil engineering data into verifiable digital assets for AI and co-pilots?
- Showcase winning end-to-end ecosystems from the ASCE 2026–2027 Tokenization Hackathon, presenting on ASCE 2027 in March, Philadelphia.
- Explore the role of tokenized data in enhancing cyber-resilience and AI validation within practical projects.
- Contribute to the strategic design of ASCE's roadmap toward a decentralized, verifiable engineering knowledge ecosystem.
- Further details regarding the competition will be formally announced.



ASCE2027

Lead: Pingbo Tang, Carnegie Mellon University · Pittsburgh, USA
Core Organization Team: Ruoxin Xiong · Kahyun Jeon · Yanyu Wang · Daniel Calabro
Contact: ptang@andrew.cmu.edu

End-to-End RDMC Lifecycle in OpenConstruction - Human Data Integrated

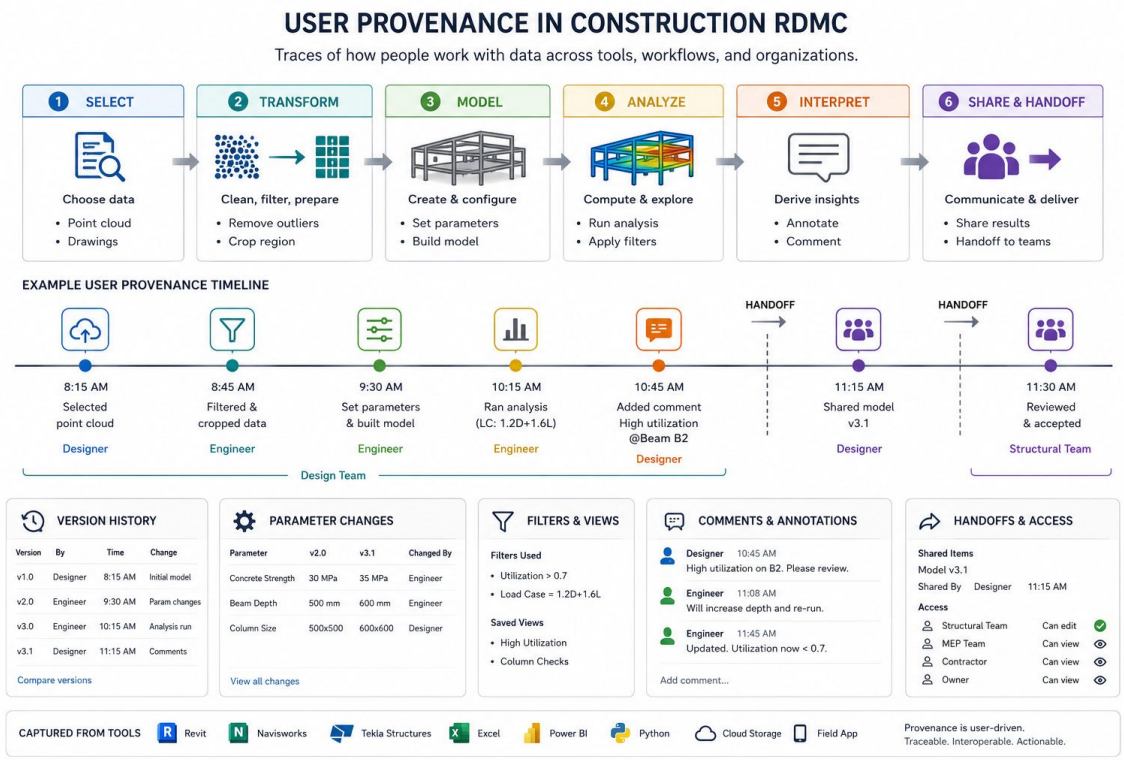


- Resources enter via images, videos, UAV data, docs, IoT streams, and code, identified through systematic review and community input.
- Submission, discussion forums, and iterative curation form the social fabric of RDMC—where human data work naturally appears.

(Xiong et al., 2026)

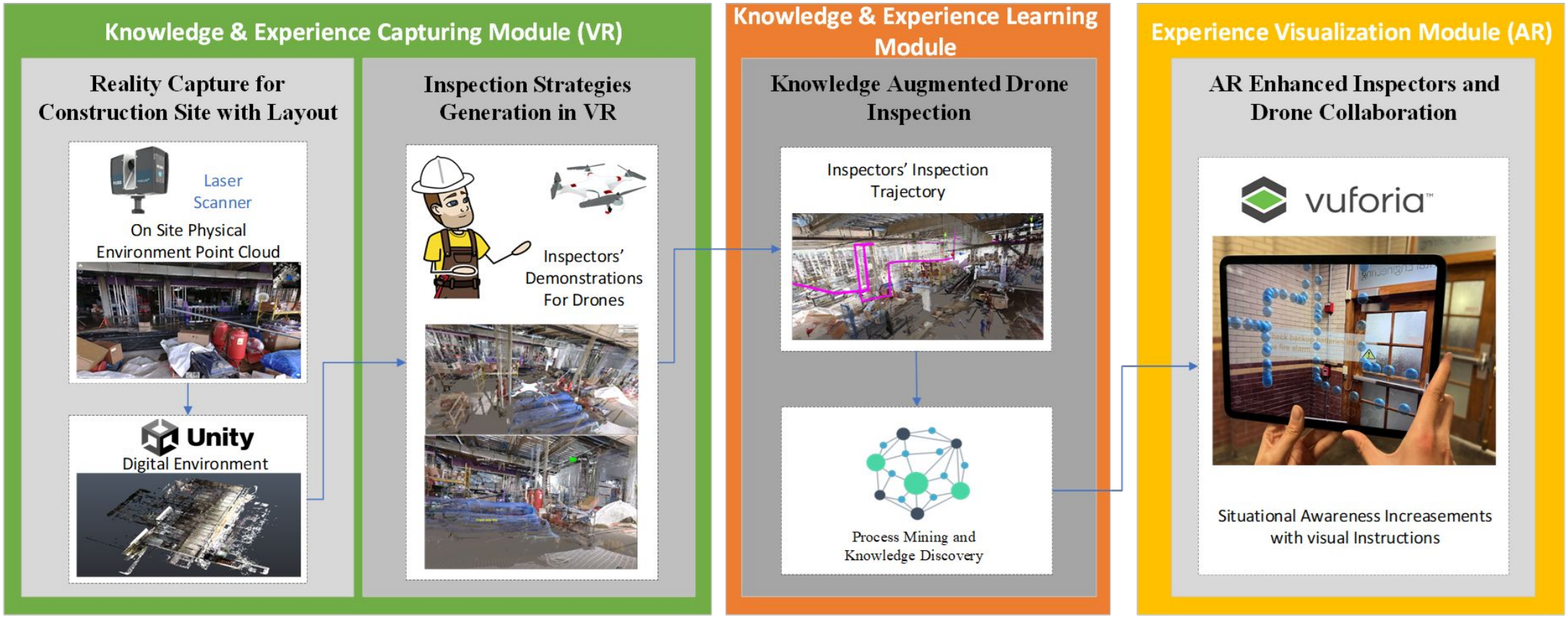
Human Data Work as a First-Class RDMC Asset

- Traces of how designers, engineers, and field workers select, transform, model, analyze, interpret, and share data inside their software tools and workflows.
- Includes version histories, parameter changes, filters, visualizations, annotations, comments, and handoffs across teams and organizations.
- These user-provenance traces complement data and computation provenance and are essential for understanding and improving RDMC in construction.



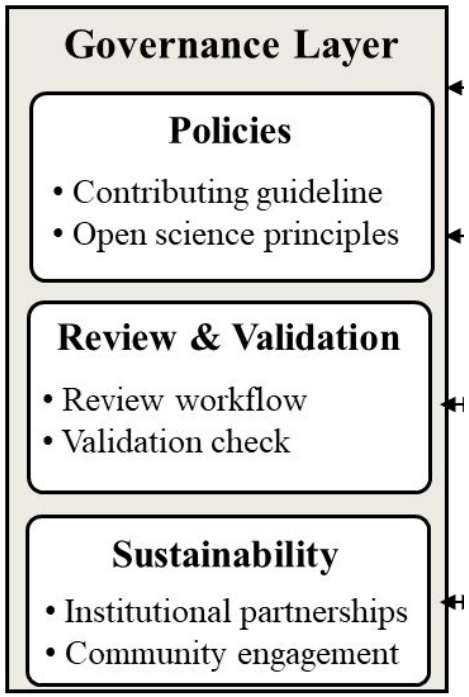
While the outcomes of our work are shared, there has been insufficient effort in documenting and sharing the underlying working processes to enhance reusability, extensibility, and generalization across diverse tasks and organizations.

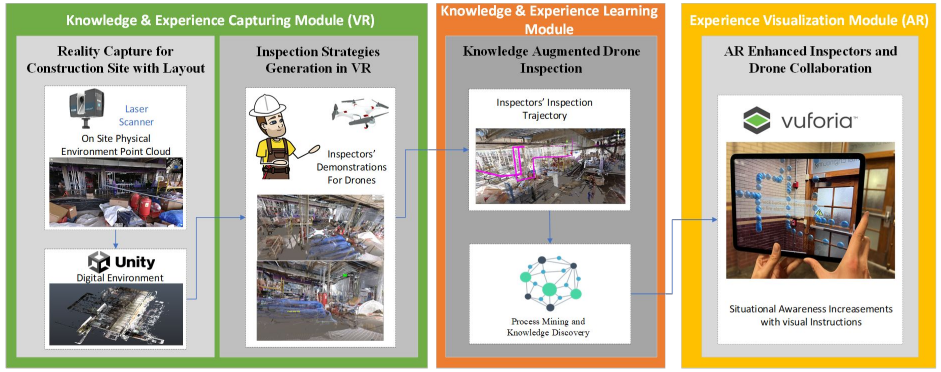
Exploring Human Data Values - Multimodal Human Data Work in Training and Operations

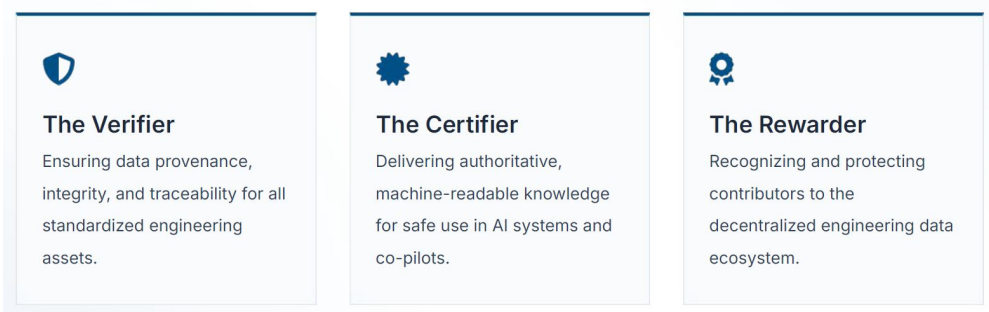


(Liu et al., 2025)

Exploring Human Data Values Across Organizations - Cybersecurity, Privacy, and Ethics

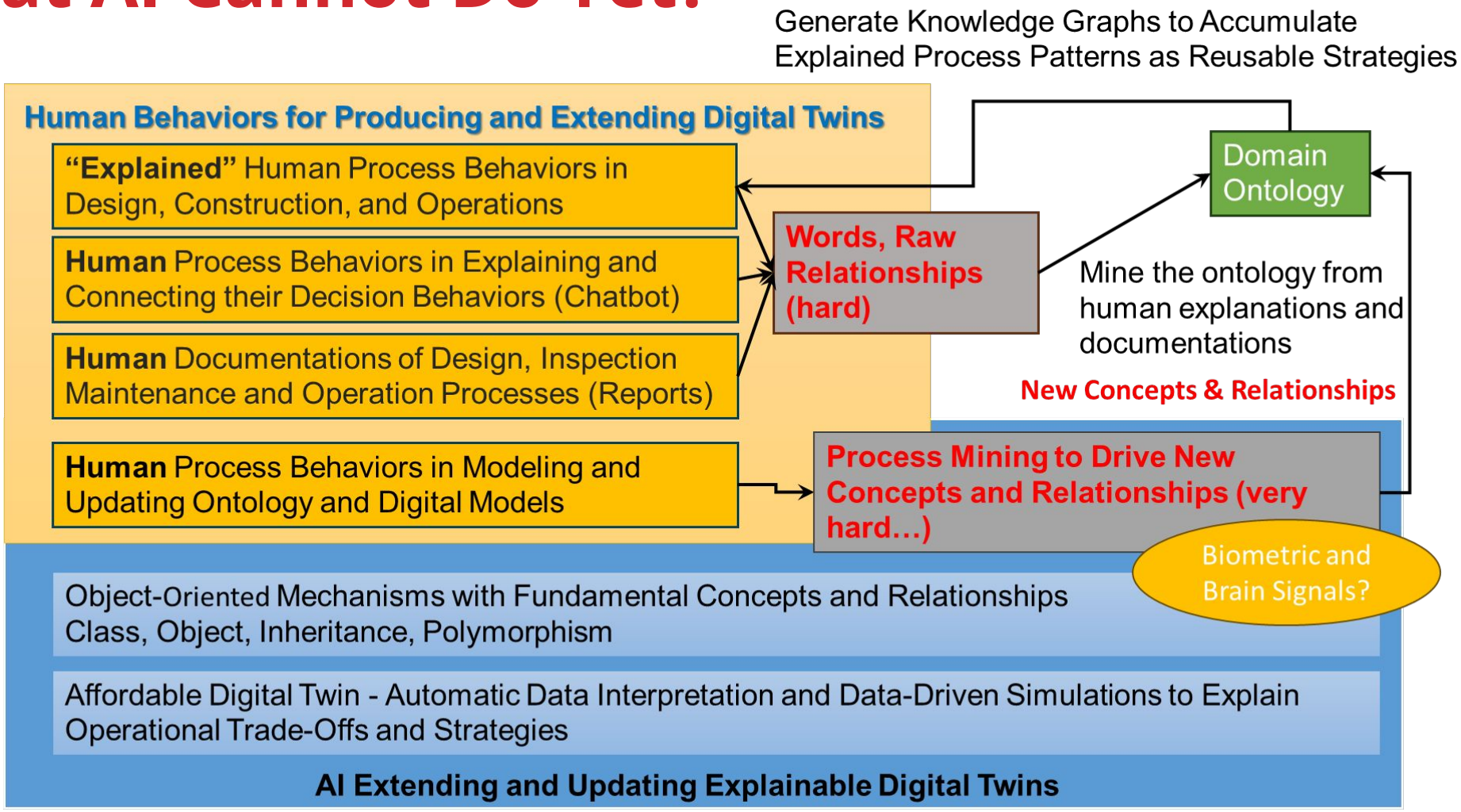






- AI captures gaze, interaction logs, decisions, and performance metrics as learners or operators use simulators and analytical tools.
- These multimodal traces show which data views, model settings, and feedback responses correspond to successful or problematic performance.
- With tokenization, such traces can be shared—under strong privacy and governance constraints—as templates of effective data work across organizations.

Human Data and Human-AI: A Broader View - What AI Cannot Do Yet?



Red— not mature yet, but we should know more about them to make "really useful" digital twins…

Final Thoughts about a Proposed Path Forward:
AI Observing Human Data Work to Evolve RDMC Principles

- **Technical:**
  - Model user-provenance traces as workflows that show which data, tools, transformations, and collaboration patterns lead to reliable, safe, and efficient outcomes.
  - Train AI agents on these tokenized traces so they can suggest next steps, flag risky or wasteful patterns, and summarize effective practices across projects and organizations.
  - Use these insights to refine data modeling, analysis, and exchange principles that respect different domain terms and processes across the design–construction–operation–maintenance lifecycle.
- **Social-technical**
  - Co-design AEC-specific cybersecurity and data-sharing guidelines for RDMC platforms and AI tools, including certification and licensing schemes appropriate for safety-critical infrastructure.
  - Develop AI policy and ethics standards for RDMC that address provenance, accountability, privacy, and equity when human data work and AI agents are tightly coupled.



ASCE
TECHNICAL ADVANCEMENT AI TASK FORCE

NSF

MICHIGAN TECH
HUSKIES

ACID
Construction Image Dataset Suite

Meet the OpenConstruction Community

Explore the researchers, practitioners, labs, and institutions growing an **open-science** ecosystem for **AI in the AEC community**.

149 CONTRIBUTORS/LABS

84 INSTITUTIONS

Contribute a Resource | **Join discussions** | **Subscribe for updates**

FUTURE WORK

- Exploring the guideline or criteria of data sharing standards, validity approaches, and protocols in civil infrastructure domain
- Exploring the cybersecurity issues in civil infrastructure data sharing
- Exploring the techniques in promoting the large-scale data sharing

REFERENCE

[1] Dean, Jared. Big data, data mining, and machine learning: value creation for business leaders and practitioners. John Wiley & Sons, 2014.

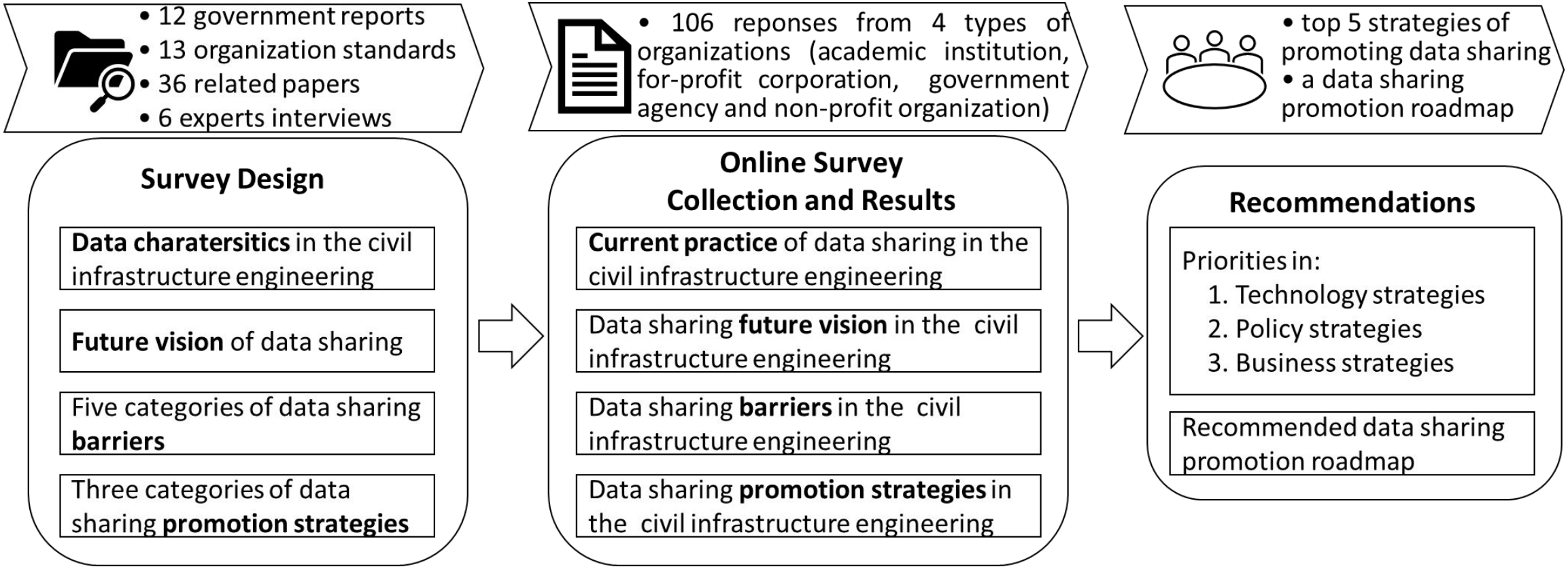
LITERATURE REVIEW

Data Sharing Related Research in Civil Engineering Domain

Reference	Data Sharing Domain	Data Sharing Current Practice	Data Sharing Barriers	Data Sharing Future Vision	Data Sharing Promotion Strategy on Technology	Data Sharing Promotion Strategy on Policy	Data Sharing Promotion Strategy on Business	Practitioners ' Priority on Promotion Strategy	Different Organizational Practitioners' Priority on Promotion Strategy
(Thu and Wehn 2016)	Hydrology	√	√	√	√	-	-	-	-
(Wang et al. 2020)	Hydrology	-	-	-	√	-	-	-	-
(Petri et al. 2017)	Construction	-	-	-	√	-	-	-	-
(Mutis and Mehraj 2022)	Construction	√	-	-	-	√	-	√	-
(Karaguzel et al. 2019)	Building Energy	-	-	-	√	-	-	-	-
(Costin and Eastman 2019)	Urban Systems	-	√	√	√	√	-	-	-
(Cooper et al. 2019)	Civil Engineering Academic Research	√	√	√	√	√	-	-	-
This research	**Civil Infrastructure Engineering**	√	√	√	√	√	√	√	√

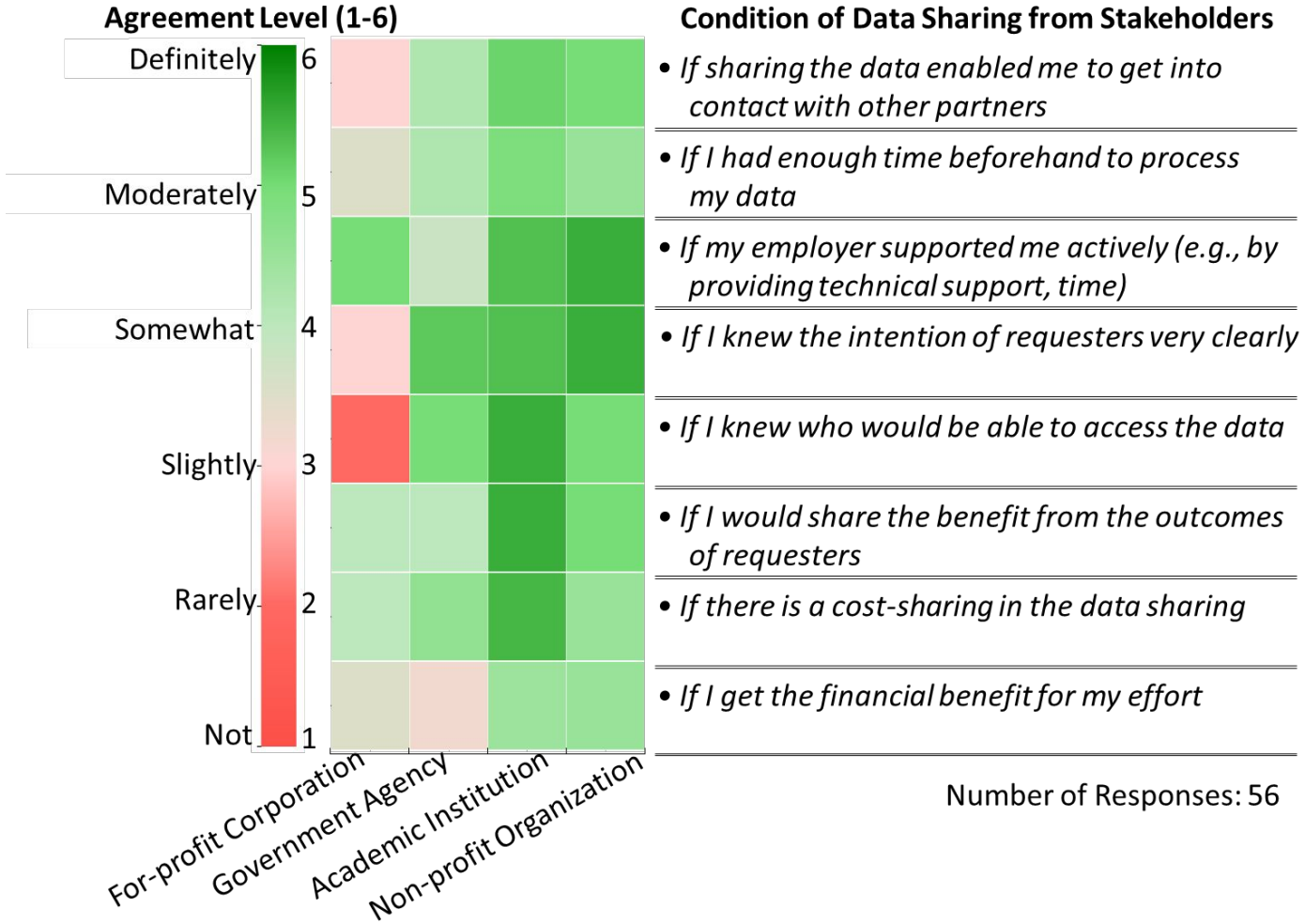
√= Discussed, − = Not Applicable

METHODOLOGY



RESULTS – FUTURE VISION

Data Sharing Conditions Perceived by Different Organizations



RESULTS – PROMOTION STRATEGIES PREFERENCE

