

Bad Data, Good Data, and the Courage to Begin

*Perfect data is not a precondition for leveraging operational AI or a digital twin.
It is a byproduct of having built these capabilities.*

By Dr. Steve Burrell

The Objection That Stops Everything

Waiting for clean data is waiting for an outcome that only the work itself produces.

Every smart campus initiative meets the same objection, usually early and usually from someone credible: *our data isn't good enough yet*. The space inventory is stale. Half the equipment records list a manufacturer and nothing else. Three different contractors named the BAS points across two decades, and nobody can say what AHU-3-RM-SP-2 actually controls. The objection is raised in good faith, and every specific claim behind it is true.

It is also the single most expensive idea in the field, because it defers indefinitely a capability that would have exposed and corrected those very defects.

Data Isn't Good or Bad. It's Good or Bad For Something

*“Bad data” is not a property of a record;
it is a mismatch between what a record supports and what someone asks of it.*

The same data field can be entirely adequate for one purpose and useless for another, and no audit of the data alone can tell you which. Consider a work order coded to a building but not a room. For chargeback allocation, it is fine — the building is the billing unit. For predicting which air handler is failing, it is worthless. Nothing about the record changed. The question changed.

This reframing matters because it dissolves the paralysis. You do not need good data. You need data adequate to a specific decision, and you can determine that in an afternoon rather than a fiscal year. *Ask what decision you are trying to improve, then ask what would have to be true of the data for it to inform that decision.* The answer is nearly always narrower than the general anxiety about data quality suggests.

Four distinct conditions hide under the single word “bad,” and they demand different responses:

- **Missing** data is honest. A null field announces itself, and systems can route around it or flag it for collection.

- **Wrong** data is dangerous precisely because it is confident. A room labeled as an office when it houses a server rack will be conditioned to the wrong setpoint by a system that has no reason to doubt itself.
- **Inconsistent** data is the most common and the most tractable. The building called SCI, Science, Science Bldg, and 0042 across four systems is not wrong anywhere — it is unreconciled everywhere, and reconciliation is mechanical work.
- **Unstructured** data is not bad at all. It is valuable data awaiting interpretation: PDF as-builts, commissioning reports, a decade of maintenance notes in free text. These are rich sources that no query can reach until someone gives them shape.

Only wrong data is genuinely hazardous, and it is typically the smallest.

Ontology: Where Meaning Lives

You will not build an ontology. Brick, Haystack, and RealEstateCore already exist, and every serious digital twin and facilities platform ships with one — the vocabulary of buildings, floors, rooms, equipment, and the relationships among them arrives in the box. This is a genuine advance, and it removes what would otherwise be a multi-year barrier to entry.

What arrives in the box is a *schema*. What does not is any statement about *your* campus.

An ontology declares that air handlers serve terminal units, which condition rooms contained by floors. Your knowledge map asserts that Air Handling Unit (AHU)-3 in the Science Building serves these eleven rooms. Two of the classrooms were converted to a lab in 2019 without a corresponding change to the space record, and the sensor in one has read ambient temps since it was knocked loose from the AHU apparatus. The schema is generic and purchasable. The instance graph is local, undocumented, and the entire object of the exercise.

This is also where the vendor's assumptions require examination. Building ontologies emerged largely from commercial real estate and research HVAC, and they model those environments well. They do not natively express instructional space bound to a term calendar, research space carrying compliance overlays, residence halls with occupancy patterns unlike any office tower, clinical space under HIPAA, or the ordinary campus reality of a single room that is simultaneously a general-purpose classroom, a department's space allocation, and a chargeback unit answering to three different owners. Extension is not a sign of poor product selection. It is where higher education's particular value gets encoded.

The payoff is the same either way: **a populated graph lets good data compensate for bad**. When entities are related rather than merely listed, a fact known in one system propagates to systems where it is missing, and a claim that is wrong in one place can be contradicted by evidence from another.

For example, suppose your space system says Room 214 is a classroom. Still, the timetable shows it has never been scheduled, the BAS shows a load profile consistent with continuous equipment operation, and the badge system records three unique credentials entering it in a year. You now have strong evidence the

space record is wrong — obtained without auditing a single record, by relating four sources that individually looked fine.

That reversal is at the heart of the argument. ***You don't populate the graph after cleaning your data. Populating it is how you discover what needs cleaning, and why it matters.*** Unrelated, an error is invisible. Related, it becomes a contradiction — and contradictions can be surfaced, ranked by consequence, and resolved in priority order.

The Case We Already Accept

*Institutional leaders sometimes treat facilities data as uniquely unreliable.
It isn't. It is simply less familiar.*

Consider HR data, which every executive in the room already trusts enough to run the institution on. Job codes drift from actual duties as roles evolve and reclassification lags. Department affiliations don't survive reorganizations cleanly. Reporting lines in the system diverge from who actually assigns the work, especially in matrixed research units. Effort allocations are estimates entered to satisfy a compliance requirement. FTE splits across cost centers are approximations negotiated between deans. Employees appear in multiple systems under variant name forms and unreconciled identifiers.

Every HR director knows this. And yet we set salary pools, model benefit costs, project retirement waves, run compensation equity studies, and satisfy federal reporting from that data — because the errors, though numerous, are mostly random rather than systematic, and analysis at scale tolerates noise that would be intolerable in an individual transaction.

The same holds across the enterprise. Student records carry stale addresses and majors changed in one system but not another. Financial data hides in miscoded chart-of-accounts strings that everyone works around. Research administration reconciles effort reporting that is honest but imprecise. The advancement database maintains multiple records for the same donor household.

The institution has always functioned on imperfect data, and it has always extracted real value from it. What we do is understand the error characteristics well enough to know which questions the data can answer and which it cannot. That competence is not new and does not need inventing for the built environment. It needs transferring.

Why the Chief Facilities Officer and CIO Must Be Co-Architects of the Data Estate

None of this happens through a handoff, and the failure modes are symmetrical.

The Chief Facilities Officer holds knowledge that exists nowhere in any system: which building's meter has read low since the 2019 retrofit, which contractor's naming conventions are trustworthy and which are fiction, what actually happens in the space the system labels as storage, which sensor has been reading

ambient temps since it was knocked loose from the AHU. This is the interpretive layer, and it is almost entirely undocumented.

The CIO brings integration architecture, semantic modeling, master data management, identity resolution, governance, and hard-won judgment about what analytics can and cannot support — plus the institutional habit of reasoning about data quality as a managed characteristic rather than a binary state.

Neither can succeed alone. The CIO who builds a model without facilities knowledge produces something technically sound and operationally naive: elegant, well-governed, and confidently wrong about what the buildings are doing. The facilities officer who pursues a digital twin as a procurement gets a vendor platform that ingests the same fragmented data and inherits every defect, now with a dashboard.

Co-architecture means **shared authorship of the mapping and the extensions** — which entities in your environment correspond to which classes in the vendor's model, what “room” means when facilities, the registrar, and space management each define it differently, and where the schema must be extended for space types the vendor never contemplated. That is not a technical decision or a facilities decision. It is a joint definitional decision with consequences for everything built afterward. Someone will make these calls regardless; if the two officers do not, an implementation consultant will, in a series of undocumented judgments during configuration.

It also means shared accountability for the operating model that follows: who owns which authoritative source, how corrections flow back to systems of record, how the twin improves rather than merely reflects the estate. And it means presenting a unified case to the cabinet, because a digital twin proposed by facilities alone reads as a capital request, and one proposed by IT alone reads as a technology project. Presented jointly, it reads as what it is — an institutional capability.

Domain Responsibility	Chief Facilities Officer (CFaO)	Chief Information / Technology Officer (CIO/CTO)
Primary Triad	SAS: Safety, Availability, Sustainability	CIA: Confidentiality, Integrity, Availability
Operational Mandate	Owns physical truth, mechanical sequences, sensor calibration, and safe operating boundaries.	Owns enterprise data architecture, API integration fabrics, identity controls, and network segmentation.
Data Contribution	Maintains point accuracy, maintenance context, and sequence logic at the physical asset.	Provides the semantic layer, secure transport, time-series storage, and cross-system access.
Co-Architect Value	Prevents IT from building abstract models detached from physical realities and safety constraints.	Prevents Facilities from purchasing isolated “Shadow OT” systems with brittle, proprietary data silos.

When the Chief Facilities Officer and CIO/CTO align under shared governance, they create an operational environment where missing or messy data is systematically identified, contextualized, and corrected. Together, they transform raw physical signals into a trusted institutional asset — turning legacy deferred maintenance and fragmented systems into a predictive, resilient, and sustainable smart campus.

Start Where the Data Is Good Enough

Good enough may be the enemy of great, but perfect is the enemy of progress.

A persistent myth in smart campus planning is that an institution must clean, audit, and integrate 100% of its physical asset data before deploying a digital twin. The opposite is true, and the sequence below is how institutions actually get there.

- **Choose a question, not a dataset.** Not “let’s model the campus” but “which buildings are conditioned outside of scheduled occupancy, and what would it save to stop?” A bounded question has bounded data requirements you can actually evaluate.
- **Scope the estate to that question.** Perhaps four buildings, HVAC only, plus the timetable and the space inventory. Everything else waits.
- **Populate relationships before perfecting records.** Getting the containment and service hierarchies right for those four buildings produces more value than a clean equipment list for all of them.
- **Let contradictions drive the cleanup queue.** The twin will tell you which errors matter, because those are the ones producing implausible results. Fix in that order, and let the discipline of correcting at the source rather than in the model become the habit.
- **Make each increment pay.** Fund the next expansion from the value of the last. This discipline separates a capability from a project, and it is only available if you start small.

Conclusion

Institutions waiting for perfect data will wait forever.

Those that start digital twin and operational AI initiatives will find that developing these capabilities is integral to achieving and maintaining data quality.

Our institutions have never had clean data and have never needed it, because we learned to understand our data well enough to ask it appropriate questions. The smart campus asks the same competence of the built environment, and it asks two executives—the CIO and the Chief Facilities Officer—to exercise it together as co-architects of a shared model of the place they both serve.

Definitions

- **Data Ontology (the shared dictionary).** A campus ontology — RealEstateCore, Brick, or Haystack — establishes a shared semantic layer. It translates proprietary vendor tags into a standard, machine-readable language defining entities, measurements, and physical boundaries. These standards are published and are embedded in most commercial digital twin platforms; the institutional work is mapping local assets onto them and extending them where higher education's space types are not represented.
- **Knowledge Map (the connected graph).** A knowledge map is the ontology instantiated against real assets, bridging cross-domain relationships across the institution. It connects a physical asset — an air handler — to its spatial context (the floors and rooms it serves), its energy profile (sub-meter data), its institutional owner (an academic department), and its human impact (scheduled class sessions).
- **Digital Twin.** A digital twin is a living, semantically structured representation of the physical campus—its buildings, systems, spaces, and the activity within them—in which static asset records, real-time operational telemetry, and institutional context are related through a shared ontology rather than merely stored alongside one another. It is not a model of a building; it is a model of a building *in operation*, continuously updated from the systems that observe it and continuously interrogated by the people and applications that act on it.

What distinguishes a twin from a dashboard, a BIM file, or a data lake is the graph. A dashboard reports values. A BIM model describes design intent, frozen at handover. A data lake accumulates. A twin asserts *relationships* — this air handler serves these rooms, which host these courses, occupied by these cohorts, metered on this circuit, maintained by this zone team — and it is those relationships that let a question asked in one domain be answered with evidence from another. The twin becomes the institution's operational memory of its own physical plant: the place where mechanical reality, spatial truth, academic use, and financial consequence are finally expressible in a single query.

For higher education, the twin is therefore neither a facilities system nor an IT system. It is the shared substrate on which operational AI, energy and carbon management, deferred maintenance strategy, space utilization, and resilience planning all depend — and the artifact that makes the Chief Facilities Officer and the CIO co-architects rather than counterparties.

Platforms such as Willow have made this practical at institutional scale, supplying the ontology, the ingestion fabric, and the time-series foundation that would otherwise take years to build — and, in doing so, moving the institutional question from *can we construct a twin* to *what will we ask it*.

References

Burrell, S. (2026). *The smart campus manifesto*. Point Peter Press.

The Smart Campus Manifesto is a guide for higher education institutions integrating AI, IoT, and digital twins, bridging IT and OT for sustainable, efficient campus operations. Written for college leaders and CIOs, it aims to improve campus experience while optimizing energy and operational costs. Available from Amazon.com.