

# The Research Data Community Has Been Solving the Wrong Problem

Why data preparation — not data curation — is the unaddressed stage of the research data chain, and why fixing it matters more than anything else the field is currently doing.

---

Russ Profant

Data Systems Specialist · 2376097 Ontario Ltd. carrying on business as PC4IT

April 2026 · [pc4it.com/research-data](https://pc4it.com/research-data)



## EXECUTIVE SUMMARY

---

The research data community has built significant infrastructure for preserving and sharing scientific data. Data Management Plans, FAIR principles, and institutional repositories represent genuine and meaningful progress. This white paper argues that these investments, while valuable, have addressed the wrong stage of the research data chain.

- Research data moves through four stages: **Data Production** → **Data Preparation** → **Data Analysis** → **Data Curation**. Of these four, only one — Data Preparation — remains entirely ad hoc, unvalidated, and methodologically invisible.
- Data production has been industrialized. Data analysis has been transformed by machine learning and computational modeling. Data curation has been professionalized through FAIR standards and institutional repositories. Data preparation still depends on research assistants with spreadsheets, learned on the job through trial and error.
- As instruments generate data faster and analysis tools require higher-quality input, the manual preparation bottleneck widens every year. The gap is structural and compounding.
- Manual data preparation introduces systematic errors that are invisible to downstream analysis. Long-term curation of incorrectly prepared data does not preserve valid science — it preserves well-archived mistakes at institutional scale.
- A Sunnybrook Health Sciences Centre neurosurgery lab recently reduced a five-month manual data preparation task to under three weeks using systematic automated processing. The time difference is not exceptional — it is representative of what the preparation gap costs every active research lab.
- The research data community should extend the methodological standards it applies to instruments and analysis to the preparation stage. This white paper offers five specific recommendations for institutions, funders, and research administrators.

## SECTION 1

# The Scientific Productivity Paradox

Scientific infrastructure has multiplied beyond recognition over the last thirty years. Instruments are more powerful by orders of magnitude. Data storage is essentially free. Compute power that would have required a supercomputer in 1990 sits in a researcher's pocket. The number of credentialed researchers has grown steadily. Global research investment has expanded significantly.

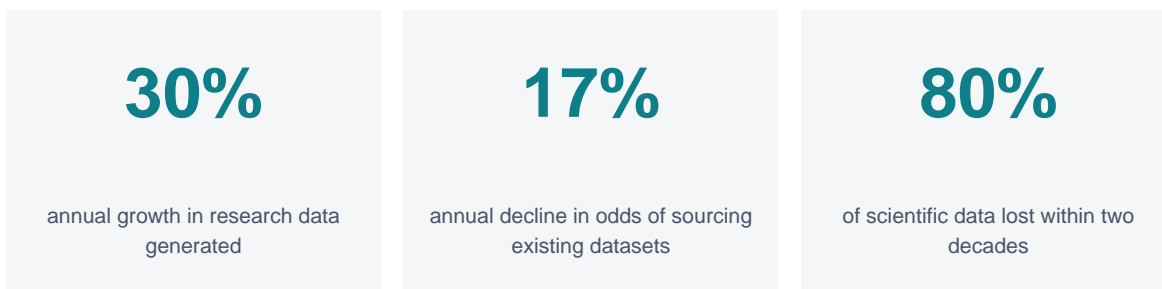
*The mobile phone — the most transformative consumer technology of our era — is built on physics and materials science developed in the 1970s. When was the last time a discovery opened an entirely new field the way quantum mechanics, the double helix, or germ theory once did?*

More money. More researchers. More instruments. The rate of genuine scientific breakthrough has slowed to a trickle. This is a complex problem with complex causes — but one contributing factor has gone almost entirely unexamined. It is not the most important cause. But it is hiding in plain sight, inside every research institution in the world, consuming an enormous quantity of the most valuable resource science has: researcher time.

## SECTION 2

# What the Field Got Right

The case for data curation is well established and correct. The scale of the data loss problem in science has been documented precisely:



*Source: American Laboratory, "What's the Real Impact of Poor Scientific Data Management?" (2014)*

The data exists. The science was done. The findings were published. And the underlying data has simply disappeared. Science advances by building on prior work. When the data behind published findings cannot be retrieved, verified, or reanalyzed, the cumulative record of science becomes a collection of claims rather

than a body of verifiable knowledge.

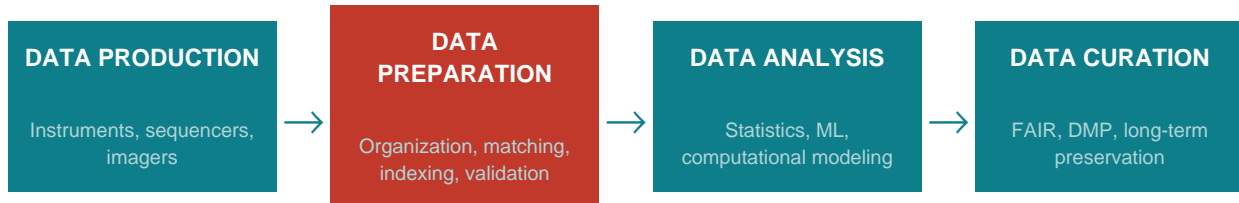
The CLIR report on data management practices among university researchers documented this problem precisely. Researchers described practices that were entirely ad hoc — learned on the job through trial and error, with no formal training and no systematic approach. None of the researchers interviewed had received formal training in data management. Most described feeling "adrift when establishing protocols" — lacking the resources to determine best practices, let alone implement them.

The field recognized this problem. It responded with Data Management Plans, FAIR principles, institutional repositories, and metadata standards. These are genuine and meaningful improvements. But the field addressed the last stage of the chain while the second stage — the one that determines the quality of everything that follows — went entirely unaddressed.

### SECTION 3

## Four Stages. One Is The Manual Middle.

Research data moves through four stages from collection to long-term use:



**DATA PREPARATION** is the only stage that remains entirely ad hoc, unvalidated, and methodologically invisible.

## SCIENCE ISN'T SLOW BECAUSE OF LACK OF TOOLS. IT'S SLOW BECAUSE OF THE MANUAL MIDDLE.

**1. DATA PRODUCTION**  
Sophisticated instruments.  
High-quality data.

**2. DATA PREPARATION – MANUAL AD HOC**  
Extracting. Cleaning. Mapping. Organizing. Enriching.  
Done by hand. In spreadsheets. Everywhere.

- THROWN OUT**  
Most data discarded
- SPOILED**  
Errors and inconsistencies
- FORGOTTEN**  
Important data never used
- MISIDENTIFIED**  
Wrong labels, wrong links

**3. DATA ANALYSIS & DISCOVERY**  
Exceptional methods.  
Powerful algorithms.

**AUTOMATE THE MIDDLE. ACCELERATE DISCOVERY.**  
We handle the data work so scientists can focus on science.

**PC4IT**  
RESEARCH DATA SERVICES  
Data ready in days, not months.

Figure 1: The Manual Middle — Data Preparation is the only stage of the research data chain that has not been modernized. It sits between sophisticated instruments and powerful analytical tools, consuming months of researcher time while introducing errors that propagate invisibly into published science.

**Data Production** has been industrialized. Sequencing a human genome costs less than a thousand dollars. MRI machines produce data at resolutions unimaginable thirty years ago. Research data is currently growing at 30% per year. Production has been transformed beyond recognition.

**Data Analysis** has kept pace. Statistical frameworks, machine learning, and computational modeling have transformed what is possible with a dataset once it is in hand. Analysis capacity has expanded dramatically and continues to grow.

**Data Curation** has been the focus of significant institutional investment over the last decade. FAIR principles, Data Management Plans, and institutional repositories represent genuine improvements in how science preserves its record. The field recognized a problem and built infrastructure to address it.

**Data Preparation** has not changed in thirty years. It is still done manually, by research assistants, learned on the job, with no documented methodology, no validation layer, and no audit trail. As instruments generate more data and analysis tools require higher quality input, this bottleneck widens every year.

**Of the four stages in the research data chain, only one is currently ad hoc, unvalidated, and methodologically invisible. It is the one that sits between every instrument and every analysis. It is the foundation on which everything else rests.**

## SECTION 4

### The Assumption Nobody Examined

---

Every data curation framework, every Data Management Plan template, every FAIR compliance checklist makes the same implicit assumption: that the data being curated was correctly prepared in the first place.

PhDonTrack, whose research data guidance reflects the standards adopted by Norwegian universities and the broader European research community, defines a Data Management Plan as describing "how research data will be managed from the start to end of a research project" — planning for future use, avoiding data loss, enabling re-use. The FAIR principles address findability, accessibility, interoperability, and reusability of data. All of these frameworks begin at the point after preparation. None address what happened before.

The CLIR study was explicit on this point: the shift to digital data collection shifted an additional burden of "labor continuity that is not readily found in a pool of transient research assistants" onto the researchers themselves — a burden for which no infrastructure or support exists. The field built a sophisticated preservation system for data that arrives at the preservation stage already potentially compromised.

## SECTION 5

### The Cost in Practice

---

A neurosurgery research lab at Sunnybrook Health Sciences Centre recently faced a data preparation task estimated at five months. Five months for a 500-patient longitudinal MRI study — approximately 2,000 paper intake forms requiring OCR extraction of patient identifiers, systematic matching to imaging files in a complex directory structure, and construction of a validated master index connecting patient identity, MRI sequence, and file location.

Five months was the accepted timeline. Nobody questioned it. It was simply how long this kind of work takes.

**The same task, approached as an engineering problem rather than a manual task, was completed in under three weeks. Every transformation was logged. A validation report flagged anomalies before they reached the analysis stage. Four and a half months of research time recovered.**

The time cost is the visible part. The invisible part is worse. When a research assistant matches 2,000 paper forms to 2,000 data folders by hand, over weeks of repetitive work, errors enter the dataset — not from incompetence, but because humans performing repetitive tasks over extended periods make systematic errors. These errors are invisible to downstream statistical methods because those methods assume the

data is what it claims to be.

**Long-term curation of incorrectly prepared data does not preserve valid science. It preserves well-archived mistakes.**

## SECTION 6

# The Standard That Is Applied Everywhere Except Here

---

Research methodology is held to rigorous standards. Protocols are documented. Instruments are validated. Controls are run. Limitations are reported. These standards exist because scientific findings need to be reproducible, auditable, and trustworthy.

Data preparation — the step that converts raw collected data into the structured dataset on which all analysis rests — is held to no methodological standard at all. It is performed ad hoc, undocumented, unvalidated, and unreproducible. A different research assistant performing the same task manually would produce a different result. That variability does not appear anywhere in the published paper.

A documented, automated data processing workflow produces the same output every time. The processing log records every transformation. A validation report identifies anomalies. The output is reproducible. This is what methodological rigour looks like applied to data preparation. The research community has simply not extended its existing standards to this step.

## SECTION 7

# Recommendations

---

The following recommendations are addressed to research institutions, funders, and research administrators. Individual researchers can act immediately by applying systematic data preparation to their next project. Institutional change requires deliberate policy.

**1****Extend methodological standards to data preparation**

Require that data preparation workflows be documented, validated, and reproducible as a condition of publication and grant compliance — the same standard applied to instruments, protocols, and analysis methods.

**2****Audit existing data management frameworks**

Review Data Management Plan templates and FAIR compliance checklists to add explicit requirements for data preparation methodology. The preparation step should not remain an assumed precondition.

**3****Recognize data preparation as a distinct research skill**

Graduate training programs should include systematic data preparation alongside existing data management and statistical training. The CLIR study found no researcher who had received formal preparation training.

**4****Pilot systematic data processing in high-volume labs**

Research institutes should identify high-volume imaging, genomic, and preclinical labs currently consuming significant researcher time on manual preparation and pilot automated processing. Measure time recovery and error reduction.

**5****Establish vendor approval pathways for data processing services**

Technology Transfer Offices and research administration should develop standard agreements for external data processing services — similar to existing frameworks for cloud storage and analytics platforms — to reduce per-project friction.

## CONCLUSION

### Completing the Picture

---

The data curation community is right that long-term preservation matters. The FAIR principles are the correct aspiration. Data Management Plans are a genuine improvement in research practice. The 80% data loss figure represents a real and serious problem that the field has correctly identified and is working to address.

But preservation of incorrectly prepared data is not a solution to the reproducibility crisis. It is an amplification of it — errors preserved at institutional scale, made findable and accessible per FAIR principles, available for reanalysis by future researchers who will inherit the preparation mistakes of their predecessors without knowing they exist.

The research data community has been solving half the problem. The missing half is not complex. It does not require new standards, new funders, or new policy frameworks. It requires applying to data preparation the same systematic, documented, validated approach that research already applies to every other methodological step.

***Data preparation is part of your methodology. The field has simply not started treating it like one.***

## REFERENCES

---

- [1] American Laboratory. (2014). *What's the Real Impact of Poor Scientific Data Management?*  
[americanlaboratory.com/Blog/156513](http://americanlaboratory.com/Blog/156513)
- [2] Jahnke, L.M. & Asher, A. (2012). *The Problem of Data: Data Management and Curation Practices Among University Researchers*. Council on Library and Information Resources. [clir.org/pubs/reports/pub154/problem-of-data/](http://clir.org/pubs/reports/pub154/problem-of-data/)
- [3] PhDonTrack. (2024). *Research Data Management*. University of Bergen / NTNU / University of Oslo / UiT.  
[phdontrack.net/good-research-practices/research-data/](http://phdontrack.net/good-research-practices/research-data/)
- [4] ScienceDirect. *Biomedical and Clinical Research Data Management*.  
[sciencedirect.com/science/article/pii/S1043661823003997](http://sciencedirect.com/science/article/pii/S1043661823003997)

---

**About PC4IT** Russ Profant is a data systems specialist with 30 years of experience building data processing infrastructure for large financial institutions including Morgan Stanley, Citibank, CIBC, RBC, and Canada Life. PC4IT offers research data processing services to academic and hospital-based research labs. [pc4it.com/research-data](http://pc4it.com/research-data) | [russ@pc4it.com](mailto:russ@pc4it.com) | 416.623.9031