



SPA MDR CDR SCE

From XPT to JSON for submissions: Are we ready for the change?

Poorna Suresh
Bill Qubeck



The Clinical Data
Science Conference

APAC 2026

19-21 February
Novotel Hyderabad International Convention Centre



About the Presenters



Poorna Suresh

Business Analyst



poorna.suresh@sycamoreinformatics.com



Bengaluru
India



<https://www.linkedin.com/in/poorna-suresh/>



Bio

Poorna Suresh is a Business Analyst at Sycamore Informatics with over 9 years of experience driving enterprise transformations. An AWS Certified Solutions Architect with a Master's in Electrical Engineering, she specializes in SCE, cloud infrastructure (AWS), and cross-functional leadership, managing complex connectivity and vendor engagements within highly regulated global environments.



Bill Qubeck

Senior Director of
Services



bill.qubeck@sycamoreinformatics.com



North Carolina
USA



<https://www.linkedin.com/in/williamqubeck/>



Bio

Over the past 30 years, Bill has led many digital transformations in clinical trial analysis and reporting, optimizing trials through advanced data strategies and emerging technologies. He has partnered with leading pharma, CROs, & technology providers to modernize clinical research, integrating new tech & next-gen analytics. With expertise in statistical computing, biometrics, and regulatory, he helps organizations harness technology to improve efficiency & drive innovation.



Agenda

- The Shift in Clinical Data Management
- JSON Core Architecture Features
- Define-XML Remains the Regulatory Source of Truth
- Parallel Submission Pipeline (PSP): Continuous Real-time Submission
- Technical and Metadata Verification
- A Quality Control Framework for Dataset-JSON
- Ensuring Data Integrity with GxP Qualification Framework
- Transforming Data into Submission Ready Analyses
- Conclusion and Future Work



The Shift in Clinical Data Management

Core Driver: Clinical data management has evolved from manual, paper-based processes to seamless, real-time digital integration, significantly improving the speed and accuracy of medical research.



The Era of Paper and Manual Records (Pre-1990s):

Reliance on paper Case Report Forms (CRFs).

Data exchange involved physically mailing or faxing large binders.

Slow, error-prone due to human transcription, and data was siloed without standardized language.



The Digital Shift: Early EDC and HL7 (1990s):

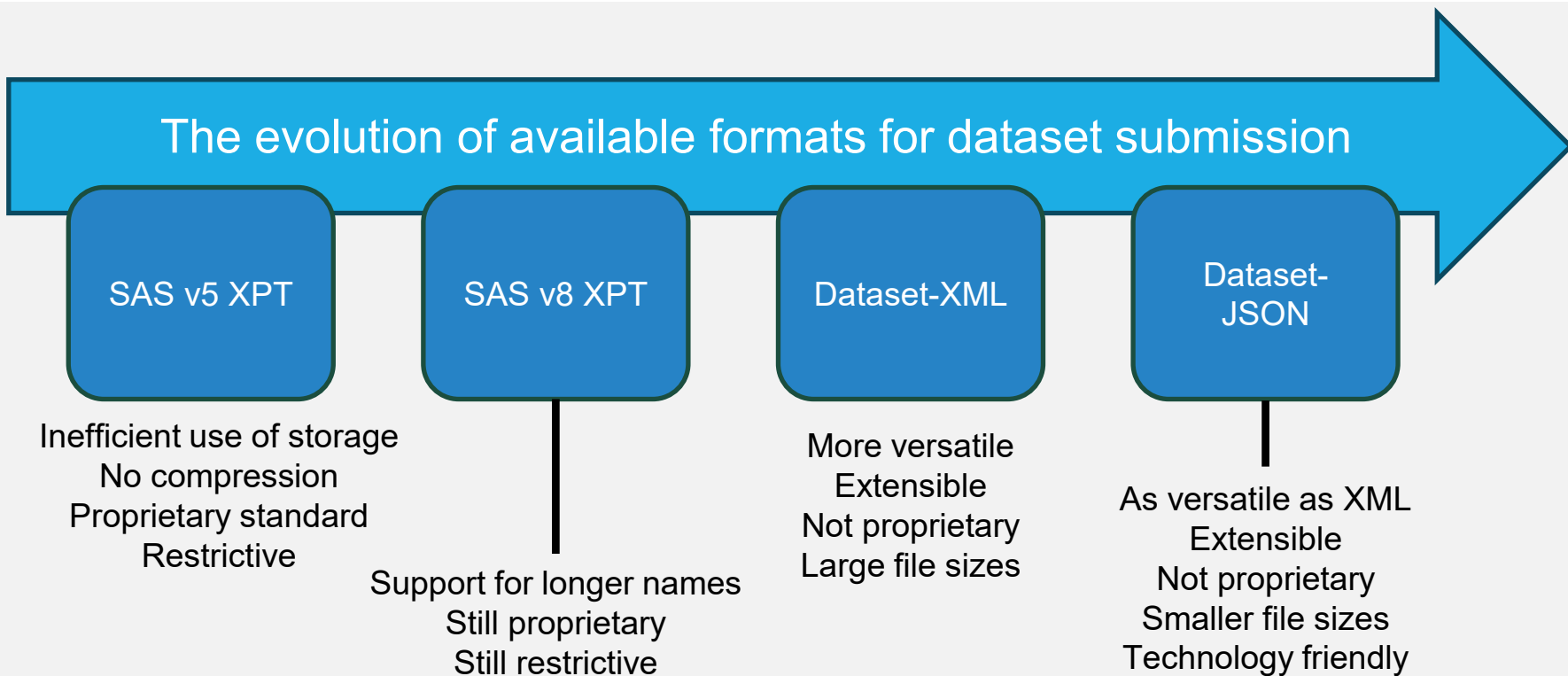
Introduction of Electronic Data Capture (EDC) systems.

Health Level Seven (HL7) emerged to standardize Electronic Health Record (EHR) exchange.

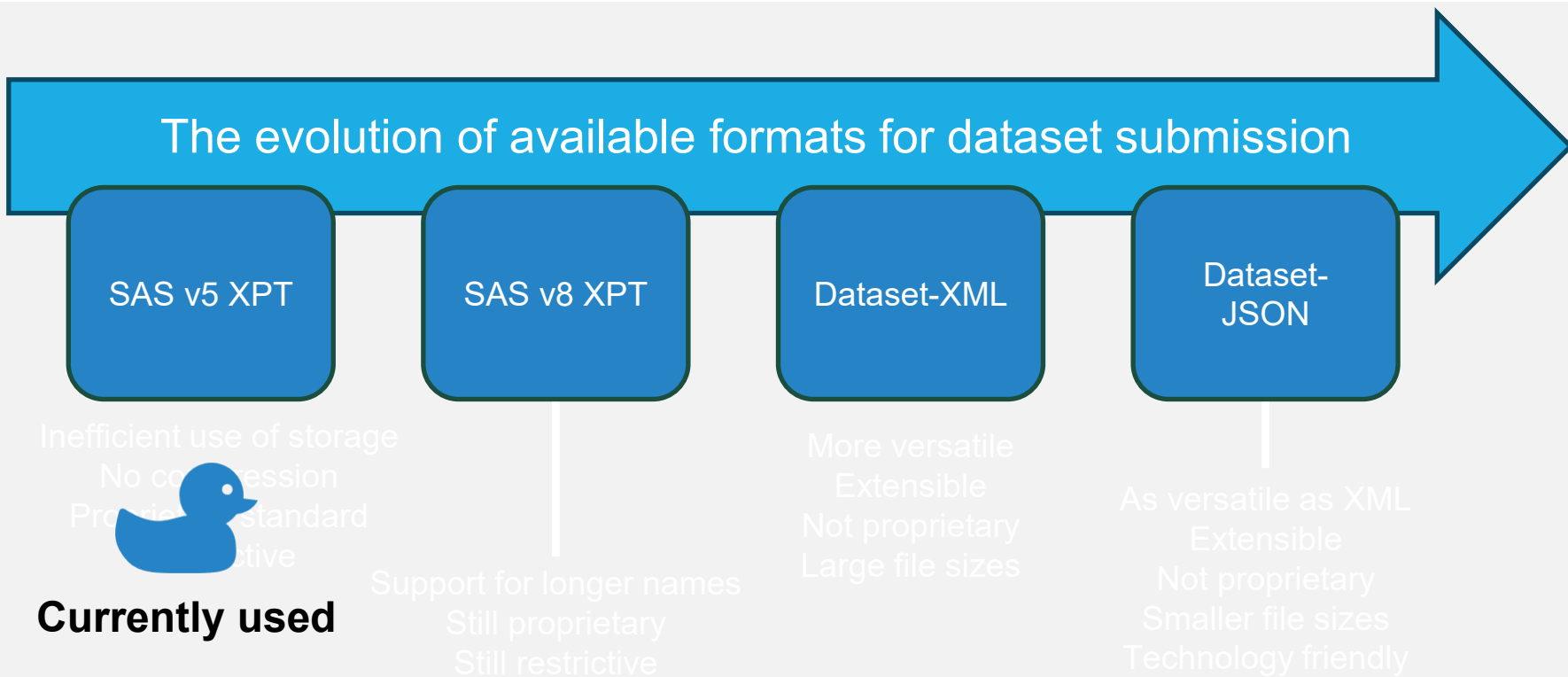
EDC systems were "siloed," using proprietary formats that limited interoperability.



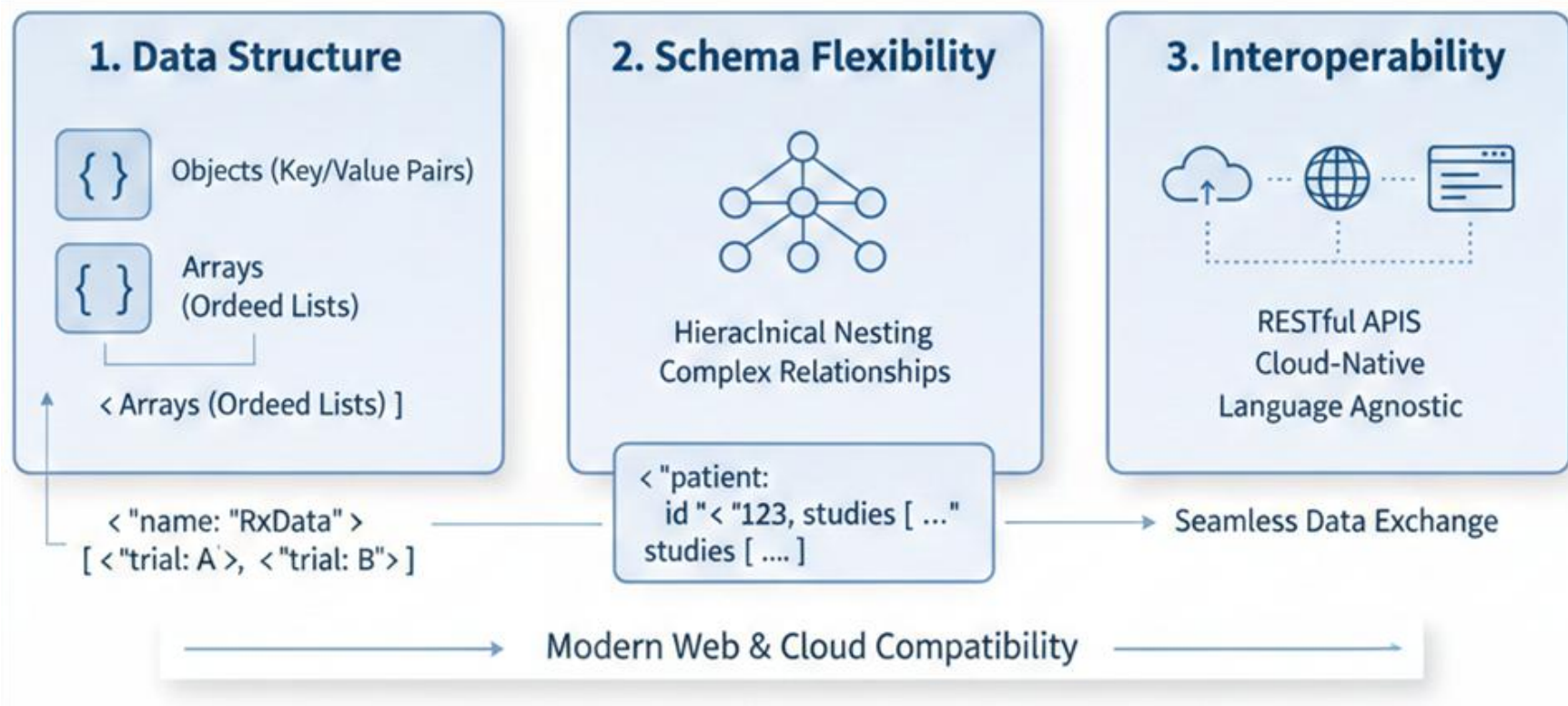
Evolution of Submission Standards



But Have We Really Evolved?



JSON Core Architecture Features



The Technical Constraints Driving the Shift to Dataset-JSON

Origin of XPT Format: A non-proprietary, open-specification binary format developed in the 1980s, originally designed for dataset transfer between mainframe and PC systems.



Key Technical Limitations:

Rigid Lengths: Var names are restricted to **8 characters**.

Limited Metadata: Var labels are limited to **40 characters**, and string values to **200 characters**.

Decoupled Metadata: Requires a separate **Define -XML** file; the data is not self-describing.

Fixed Structure: **Tabular only**, which struggles to handle complex, hierarchical data (e.g., genomics, sensor data).

Encoding: Limited to US-ASCII, preventing the use of international characters (UTF-8).

Readability: Not human-readable requires specific software (like SAS, R, or Python) to open and interpret.



Define-XML Remains the Regulatory Source of Truth

- **Core Principle:** While Dataset-JSON is the data *transport* format, **Define- XML** remains the **authoritative source of truth** for regulatory submissions. DSJ files must be validated against it.
- **The Metadata Divide:** DSJ focuses on what is needed to process the file; Define-XML holds the comprehensive study-level context.
- **Key Linkage:** studyOID & metaDataVersionOID in the Dataset-JSON act as foreign keys that must match the identifiers in the associated Define-XML file, ensuring correct context for reviewers.

Feature	Contained in Dataset-JSON?	Contained in Define-XML?
Variable Names & Labels	Yes	Yes
Data Types	Yes (JSON-specific)	Yes (Standard-specific)
Codelists / Controlled Terminology	No	Yes
Value-Level Metadata	No	Yes
Origin / Derivations	No	Yes
Method / Comments	No	Yes



Metadata Hierarchy and Strategic Value in Regulatory Submissions

Metadata Hierarchy:

- **Define-XML:** Remains the **Authoritative Source of Truth** for comprehensive metadata (derivations, codelists, comments).
- **Dataset-JSON:** Acts as the **Technical Transport Format** for the data payload. It embeds essential structural metadata (item names, types) but relies on Define-XML for external context.

Strategic Value in Data Architecture:

- **Efficiency and Scalability:** Minimalistic syntax ensures smaller data packets, optimizing for **cloud-native environments** and high-throughput data exchange.
- **Regulatory Streamlining:** Facilitates **self-describing data structures** and is optimized for the cloud, paving the way for faster data transfers, real-time monitoring, and rolling reviews with agencies like the FDA.
- **Digital Transformation:** The shift is foundational to modernizing the clinical data pipeline, moving away from proprietary dependencies to an open, flexible data ecosystem.

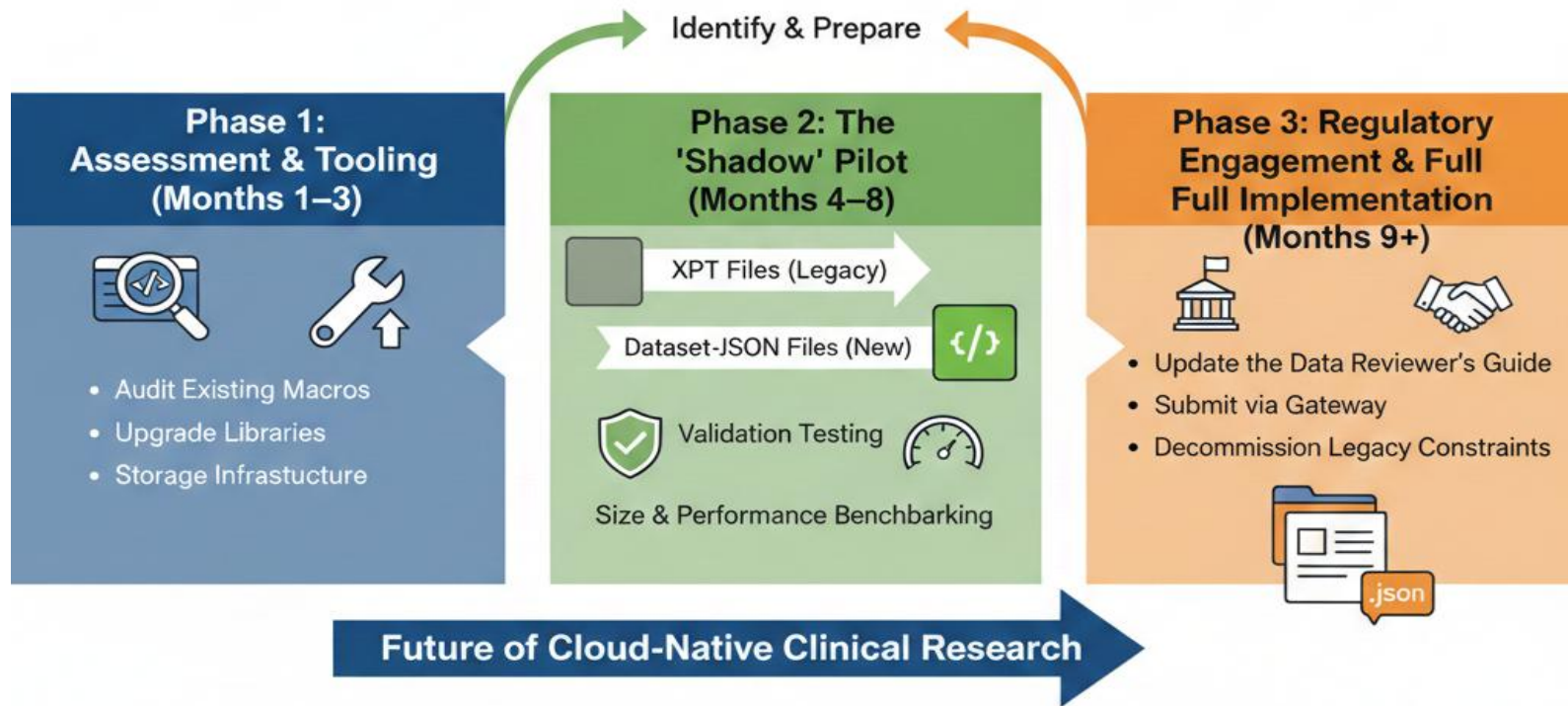


The Parallel Submission Pipeline (PSP): Continuous, Real-Time Submission

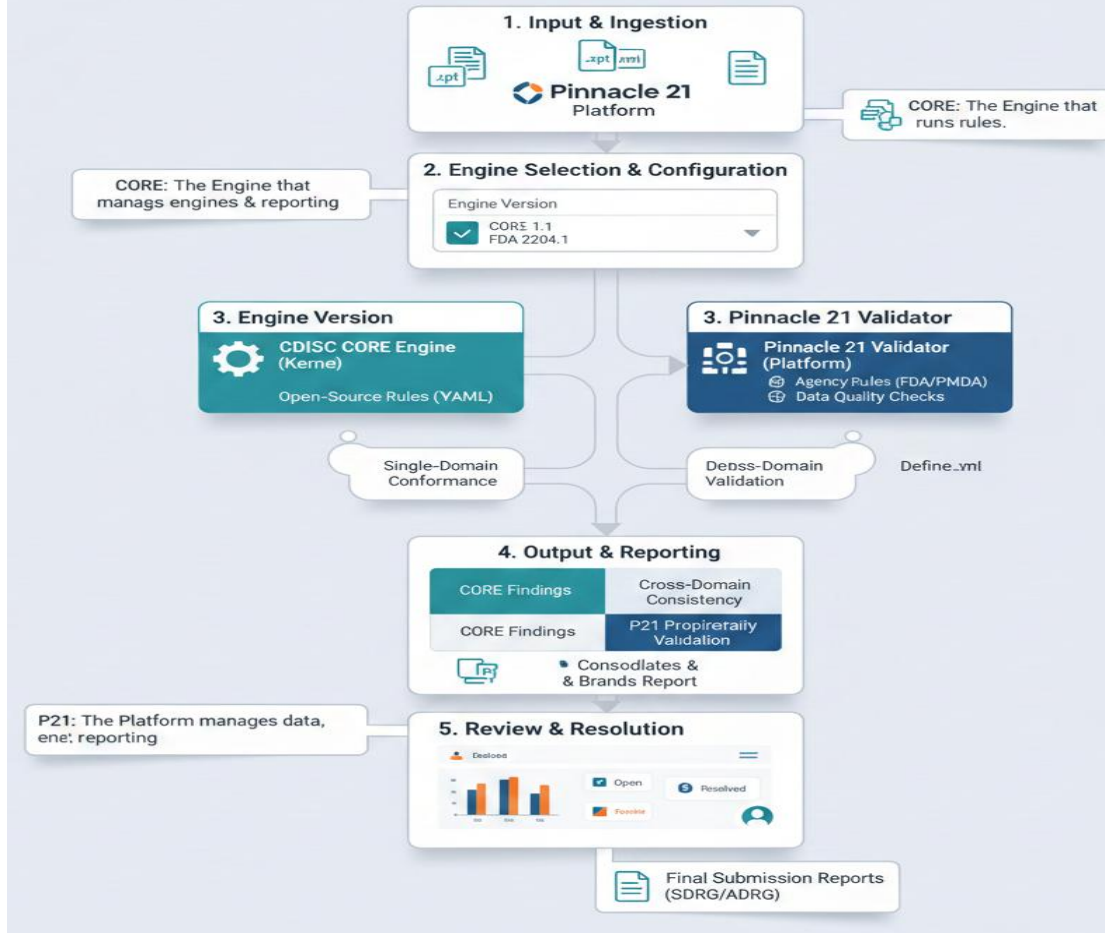
- **Three Concurrent Workstreams:**
 - **Continuous Data Stream:** Clinical data is ingested, cleaned, and SDTM-converted in monthly or event-driven batches, rather than waiting for the end of the study.
 - **Iterative Shelling & Authoring:** Medical writers begin drafting key regulatory documents using interim or "dry-run" data, instead of waiting for the final Tables, Listings, and Figures (TLFs).
 - **Simultaneous Global Filings:** Maintains a Global Master Dossier to enable near-simultaneous submissions across multiple Health Authorities (HAs).



Strategic Timeline for Dataset-JSON Adoption (12-24 Months)



Clinical Data Validation Flow: CORE & Pinnacle 21 Integration



Technical and Metadata Verification Steps for Data Integrity

- **Technical Verification Steps:**

- **Schema Validation:** Use a JSON Schema validator to ensure the file complies with the official CDISC structure (detects missing commas, incorrect metadata types).
- **Encoding Check:** Verify that the file is strictly encoded in **UTF-8** (a major advantage over XPT's ASCII limit).
- **Precision & Data Types:** Ensure numeric precision is maintained (especially for floating-point numbers) and dates follow the **ISO 8601** format.
- **File Integrity:** For large datasets (5GB+), check for file truncation; the use of **NDJSON** (Newline Delimited JSON) is being introduced to facilitate streaming and QC.

- **Verification of Metadata Linkage:**

- The **Define-XML** file remains the authoritative source of truth.
- **OID Matching:** The itemGroupOID in the JSON must exactly match the ItemGroupDef OID in the Define-XML.
- **Variable Order:** The column order in the JSON's items array must match the order defined in the Define-XML to prevent mapping errors during review.



Comparing Formats and Avoiding Common Pitfalls

Comparison Techniques

dsjconvert (CLI Tool): A specialized Python tool for bidirectional conversion between XPT and Dataset-JSON v1.1, enabling comparison using standard JSON diff tools.

Lex Jansen's SAS Macros: Open-source macros for reading Dataset-JSON directly into SAS datasets, allowing data comparison using standard tools like PROC COMPARE.

Direct Python Comparison: Utilizing pandas and libraries like datacompy to read both XPT and JSON into DataFrames for a side-by-side data value comparison.



Common Pitfalls and False Positives

Mismatch in OIDs: The most critical error, where the header OID in the JSON doesn't align with the Define-XML.

Null Handling: Incorrect representation of missing data (e.g., using an empty string "" instead of the JSON null).

Variable Lengths: Differences caused by XPT's rigid 8-character variable name and 40-character label limits, which JSON does not have.

Character Encoding: Issues arising from XPT being US-ASCII and JSON being UTF-8.



A Quality Control (QC) Framework for Dataset-JSON

- QC for Dataset-JSON is essential to ensure the data is fit for regulatory use (e.g., FDA/EMA) and is typically organized into three distinct layers.
- Tools for QC:
 - **CDISC CORE:** The primary open-source engine for executing CDISC conformance rules.
 - **Pinnacle 21/Beacon:** Commercial validators with native support to read and validate Dataset-JSON.

Layer	Focus Area	Key Verification Points
Technical / Schema	JSON Structure	Compliance with the CDISC Dataset-JSON Schema (e.g., v1.1). Checks for required fields like itemGroupData, creationDateTime, and datasetJSONVersion.
Metadata Integrity	Define-XML Alignment	Ensuring the OIDs (Object Identifiers) and metadata in the JSON file match exactly with the corresponding Define-XML file.
Content / Conformance	CDISC Standards	Traditional SDTM/ADaM/SEND conformance rules (e.g., variable lengths, controlled terminology, and cross-domain consistency).



Ecosystem and Tooling for Dataset-JSON Generation and QC



R Ecosystem: The Lead in Innovation

{datasetjson}: The industry standard R package for wrapping data frames with CDISC metadata and handling complex type conversions (e.g., R factors/dates to ISO 8601 strings).



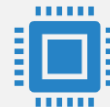
Python Ecosystem: The Scalability Choice

dsjconvert (CDISC Contributors): A dedicated CLI tool for bidirectional conversion between XPT and Dataset-JSON. Favored for high-performance processing of large datasets (Big Data).



SAS Ecosystem: The Legacy Bridge

Lex Jansen Macros: Widely adopted open-source macros to automate wrapping SAS datasets into the complex Dataset-JSON structure using PROC JSON and DATA steps.



Tool-Based Comparison:

Visual Diff tools (like Beyond Compare) and **Command Line** tools (like jq and csdiff) are also recommended.



Ensuring Data Integrity with the GxP Qualification Framework

GxP Approach: Ensuring Validation and Control

Core Objective: Ensure every system, equipment, and process is **Fit for Intended Use** and maintains a **State of Control** throughout its lifecycle in clinical research.

Change Control (Maintaining the State)

- A formal process to assess and approve any proposed changes (software, hardware, process) that could impact the system's validated state or data integrity.
- Key Requirements: Impact Assessment, Full **Traceability (ALCOA+)**, Secure Audit Trails, and Approval by Quality Assurance (QA).

Emerging Trend: Shifting to a **Quality Risk Management** approach where the level of qualification (**Tiered Validation**) is based on the risk the data poses.



GxP Qualification framework for Statistical Computing Environment

Ensure data integrity, traceability, reproducibility of Clinical outputs (TLF) in alignment with 21 CFR Part 11 and GAMP standards

Item	Requirement	Proof of Compliance
ALCOA+	Data must be Attributable, Legible, Contemporaneous, Original and Accurate	System-generated time-stamped audit trails
Version Control	Every script must be versioned	Integration with Git/SVN within the SCE
Package Vetting	Open-source code must be checked	Risk Assessment Report for each R/Python packages



Transforming Data into Submission Ready Analyses

Data Validation

- Ensure **accuracy**, completeness, and consistency of data
- Implement **automated** validation tools to detect and correct anomalies
- Establish **regular** calibration processes for long-term data reliability

Reproducible Results

- Implement **lineage** and traceability across acquisition, processing, and analysis with built in version control
- Adopt standardized **data formats** & analysis programs to enhance study consistency
- Conduct regular **audits** to maintain protocol adherence and data integrity

Standardized Reporting

- **Automate** report generation (global programs, tools)
- Define **clear and consistent** data handling, statistical analyses, and visualizations
- Adhere to regulatory reporting **standards**, including data anonymization & ethical considerations

Regulatory & Compliance

- Good Clinical Practice (GCP) & FDA 21 CFR Part 11 for **data integrity**
- Implement robust **access controls**, audit logs, and data encryption
- Conduct internal **audits** & regulatory consulting to identify and mitigate risks



Dataset-JSON v1.1 and Beyond: Handling Big Data

- The release of Dataset-JSON v1.1 addresses critical findings from industry pilots to ensure the standard is robust for global and "Big Data" use.
- **Key Technical Advancements:**
 - **Support for NDJSON (Newline Delimited JSON):**
 - **Purpose:** To handle extremely large datasets (e.g., high-frequency sensor data, large genomics files).
 - **Benefit:** Allows for data **streaming line-by-line**, which prevents memory crashes associated with loading multi-gigabyte monolithic JSON files.
 - **Precision and Data Types:**
 - A new **decimal data type** has been introduced.
 - **Mechanism:** Numbers are stored as strings to prevent the **rounding errors** common in standard JSON floating-point libraries.



The Future: Beyond 'Same Content, Different Suitcase'

- **Current Focus:** Initial use cases focused on simply mapping existing SDTM/ADaM data into JSON (the "same content, different suitcase" approach).
- **The Future Strategy:** The focus will shift to leveraging the inherent power of JSON's nested and flexible structure.
- **Upcoming Capabilities:**
 - **Integration of New Data Types:** Moving to natively support complex data such as high-frequency sensor and genomics data.
 - **Directly Embedded Audit Trails:** Leveraging JSON's nesting to include **Audit Trails** directly within the submission package.
 - **Source Data Inclusion:** Integrating **Source Data** (e.g., EHR or Lab transfers) directly into the submission package.
- **Long-Term Vision:** Dataset-JSON ensures clinical assets are **future-proofed**, remain accessible, and are actionable for years to come, completing the digital transformation of clinical research.



Q & A

Connect with the Authors!

Bill Qubeck



Poorna Suresh



Thank you!



phuse·connect
connect · share · advance

The Clinical Data
Science Conference

APAC 2026

19–21 February
Novotel Hyderabad International Convention Centre