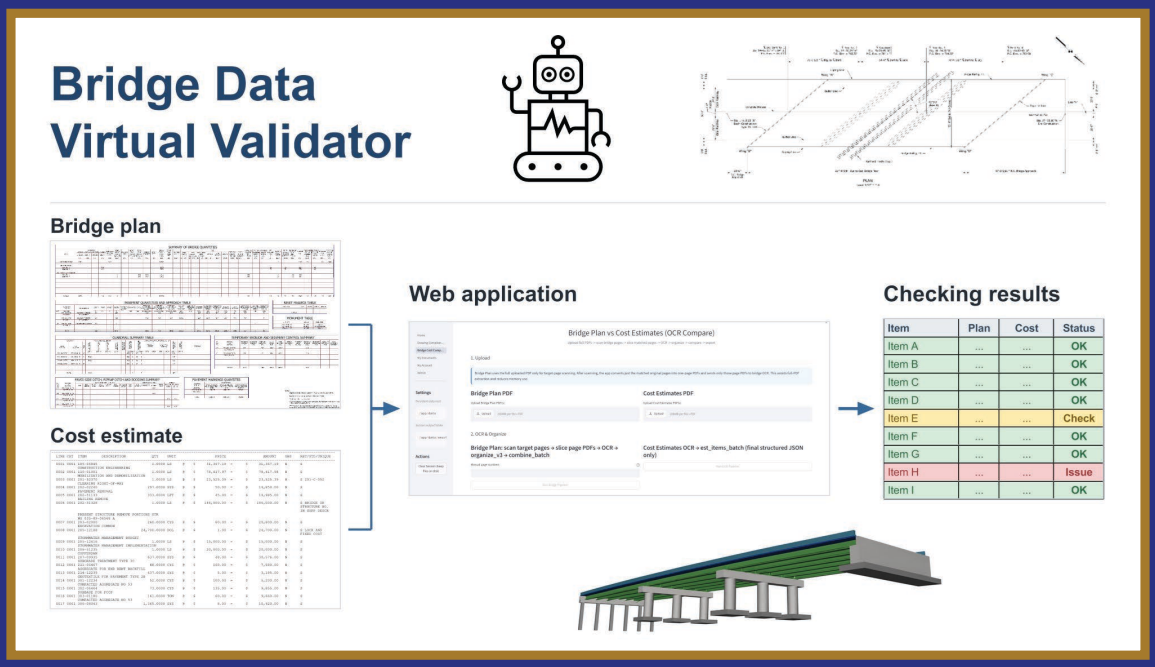
JOINT TRANSPORTATION
RESEARCH PROGRAM
INDIANA DEPARTMENT OF TRANSPORTATION
AND PURDUE UNIVERSITY

Bridge Data Virtual Validator



Jiansong Zhang, Bo Pang, Meta Soy, Bernard Nartey, and Yunfeng Chen

SPR-4927 • Report Number: FHWA/IN/JTRP-2026/23 • DOI: 10.5703/1288284318701

RECOMMENDED CITATION

Zhang, J., Pang, B., Soy, M., Nartey, B., & Chen, Y. (2026). *Bridge data virtual validator* (Joint Transportation Research Program Publication No. FHWA/IN/JTRP-2026/23). West Lafayette, IN: Purdue University. https://doi.org/10.5703/1288284318701

AUTHORS

Jiansong Zhang, PhD
Purdue University
(765) 494-1574
zhan3062@purdue.edu
Corresponding Author

Bo Pang
Purdue University

Meta Soy
Purdue University

Bernard Nartey
Purdue University

Yunfeng Chen, PhD
Purdue University

ACKNOWLEDGMENTS

This study was conducted as part of the Joint Transportation Research Program SPR-4927 project. The authors gratefully acknowledge the members of the Indiana Department of Transportation Study Advisory Council including Jennifer Hart, Jim Lesh, Mark Joseph, Maxwell Black, Prince Baah, and Stephanie Wagner for their valuable guidance and constructive feedback throughout the study.

JOINT TRANSPORTATION RESEARCH PROGRAM

The Joint Transportation Research Program serves as a vehicle for INDOT collaboration with higher education institutions and industry in Indiana to facilitate innovation that results in continuous improvement in the planning, design, construction, operation, management and economic efficiency of the Indiana transportation infrastructure. Learn more at engineering.purdue.edu/JTRP.

Published reports of the Joint Transportation Research Program are available at docs.lib.purdue.edu/jtrp/.

NOTICE

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views and policies of the Indiana Department of Transportation or the Federal Highway Administration. The report does not constitute a standard, specification, or regulation.

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. FHWA/IN/JTRP-2026/23	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Bridge data virtual validator		**5. Report Date** August 18, 2026 **6. Performing Organization Code**
7. Author(s) Jiansong Zhang, PhD (https://orcid.org/0000-0001-5225-5943) Bo Pang (https://orcid.org/0009-0003-1177-292X) Meta Soy (https://orcid.org/0000-0003-4826-4964) Bernard Nartey Yunfeng Chen, PhD (https://orcid.org/0000-0002-0108-8484)		**8. Performing Organization Report No.** FHWA/IN/JTRP-2026/23
9. Performing Organization Name and Address Joint Transportation Research Program Hall for Discovery and Learning Research (DLR), Suite 204 207 S. Martin Jischke Drive West Lafayette, IN 47907		**10. Work Unit No.** **11. Contract or Grant No.** SPR-4927
12. Sponsoring Agency Name and Address Indiana Department of Transportation (SPR) State Office Building 100 North Senate Avenue Indianapolis, IN 46204		**13. Type of Report and Period Covered** Final Report **14. Sponsoring Agency Code**
15. Supplementary Notes Conducted in cooperation with the U.S. Department of Transportation, Federal Highway Administration.		

16. Abstract

This report presents Bridge Data Virtual Validator, which is a prototype web-based tool developed to support automated quality assurance (QA) of INDOT bridge project deliverables. The project focuses on Final Tracings documents, where important bridge data are distributed across bridge plans, cost estimates, and other documents/elements. Two validation components were developed. The first component supports bridge plan format compliance checking by using optical character recognition (OCR), computer vision, and rule-based reasoning to identify missing or inconsistent letting control information such as file numbers, dates, stamps, signatures, and sheet numbers. The second component performs comparative checking of pay items between bridge plans and cost estimates with OCR-based information extraction, rule-based structuring, hybrid matching, and deterministic consistency checking. The comparative-checking algorithms were refined through a data-driven development process and with manually reviewed discrepancy records. Both components preserve source evidence and generate traceable reports so that reviewers can locate, verify, and resolve potential issues. The project demonstrates how existing PDF-based bridge deliverables can be converted into structured validation records without changing current submission formats, providing a practical foundation for future pilot deployment, reviewer feedback integration, and broader bridge data QA automation.

17. Key Words	18. Distribution Statement
bridge data validation, final tracings, OCR, bridge plans, cost estimates, pay item consistency, document QA, web application	No restrictions. This document is available through the National Technical Information Service, Springfield, VA 22161.

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	30	

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

EXECUTIVE SUMMARY

Introduction

Bridge project delivery depends on a large set of contract deliverables that describe the same project from administrative, design, construction, cost, and asset-management perspectives. In Indiana Department of Transportation (INDOT) Final Tracings Submittals, bridge plans, cost estimates, quantity information, Bridge Rating (BrR)–related information, and other supporting files are expected to remain complete, internally consistent, and traceable (INDOT, 2024). These deliverables are not only records for letting and construction, but also data sources that may be reused by downstream offices for inspection, asset management, maintenance, and future digital delivery workflows.

Bridge data validation has become increasingly important as transportation agencies move toward more structured bridge data and digital delivery. At the national level, bridge data validation emphasizes whether submitted data satisfy requirements such as data type, valid value, format, uniqueness, and presence (Federal Highway Administration [FHWA], 2025). State-level bridge quality control workflows also recognize the need to check bridge contract and drawings, design computations, and estimates before project delivery decisions are made (New York State Department of Transportation [NYSDOT], 2021). These efforts indicate that bridge deliverable quality assurance (QA) should address not only whether documents are submitted, but also whether the information contained in those documents is complete, valid, and consistent.

Although Final Tracings documents are submitted electronically, many required data elements remain embedded in Portable Document Format (PDF) sheets, tables, title blocks, signatures, stamps, estimate records, and model-related files. As a result, electronic submission does not automatically ensure that the submitted information is ready for automated validation or downstream use. Prior research and INDOT-related projects have shown that optical character recognition (OCR), natural language processing, machine learning, computer vision, and rule-based reasoning can extract selected bridge information from plans, drawings, tables, and text blocks (Shafei, 2024; Zhang et al., 2026). However, these technologies must be adapted to INDOT Final Tracings workflows, where validation requires cross checking information across existing deliverables rather than assuming that all information is already available in a structured database or model.

This project was initiated to develop an automated Bridge Data Virtual Validator to support QA of bridge project deliverables. The project focuses on Final Tracings documents, extracts selected information from existing files, compares related data elements across deliverables, and reports potential noncompliance items or discrepancies with evidence for engineering review. The validator is intended to assist reviewers in identifying redundant or inconsistent data-entry locations, and to in turn support future implementation pathways such as Indiana Total Assets Management Systems (iTAMS) data entry and bridge birth certificate workflows.

The specific objectives of this study are to:

1. Identify bridge data items in INDOT Final Tracings Submittals that can or should be cross validated across bridge plans, cost estimates, and other related project deliverables.
2. Develop explainable data validation algorithms that use OCR, natural language processing, machine learning, computer vision, and rule-based reasoning where appropriate to extract, structure, and compare selected bridge data items.
3. Develop and evaluate initial validation modules for bridge deliverable QA, including bridge plan format compliance checking and cross-document bridge data comparative checking.
4. Generate reviewable reports that preserve source evidence and support human-in-the-loop engineering decisions rather than automatic document approval.
5. Provide recommendations and implementation guidance for INDOT to adopt the Bridge Data Virtual Validator and improve Final Tracings data collection, validation, and downstream use.

Findings

- A web application prototype of the Bridge Data Virtual Validator was developed with two main functions: bridge plan format compliance checking and bridge data comparative checking.
- Bridge deliverable data objects suitable for automated validation were identified and tested using real INDOT Final Tracings data. The current implementation focuses on bridge plan control information and pay-item consistency between bridge plans and cost estimates.
- For bridge plan format compliance checking, an OCR-, computer vision-, and rule-based method was developed to check file numbers, signatures, stamps, dates, and page numbering and to generate traceable non-compliance reports. The method was applied to 78 INDOT bridge projects; the compliance checking system successfully detected noncompliance cases with 100% recall, meaning all noncompliance cases were detected, demonstrating its effectiveness.
- For cross-document bridge data comparative checking, an OCR-based and hybrid-matching method was developed to detect pay-item discrepancies between bridge plans and cost estimates. The method was applied to 50 INDOT bridge projects and generated traceable reports for engineering review.

Implementation

- The Bridge Data Virtual Validator is recommended as a reviewer-assistance tool. It can prescreen Final Tracings deliverables, identify potential noncompliance items and discrepancies, preserve source evidence, and support reviewer confirmation before any document correction or approval decision.
- The bridge plan format compliance checking and cross-document comparative checking functions should be maintained as complementary modules within one project-level review application. Compliance flags and pay-item discrepancy flags should remain separated in the interface and reports so that reviewers can assign, interpret, and resolve each type of issue appropriately.

- Reviewer decision fields should be added to the application so that each reported item can be marked as a confirmed issue, false alarm, resolved issue, or item requiring follow-up. These decisions should be stored with source evidence and reviewer comments.
- The Bridge Data Virtual Validator should be piloted on additional Final Tracings projects representing different bridge types, project types, consultants, document formats, and scan qualities. Pilot review should track not only algorithm performance, but also reviewer effort, confirmed issue types, false alarm patterns, and the time required to resolve each report.
- Target-page title lists, OCR correction rules, normalization dictionaries, matching rules, and compliance definitions should be maintained as versioned resources so that the tool can be updated as more projects are reviewed and documentation practices evolve.
- Future implementation could also expand the validation scope. Quantity calculations and BrR-related information could be evaluated as future data sources once the plan and cost estimate workflow is stable. The longer-term implementation pathway should connect validated Final Tracings records to iTAMS and bridge birth certificate workflows after data ownership, approval status, and transfer requirements are defined.

CONTENTS

1. INTRODUCTION . . . 9
2. PROBLEM STATEMENT . . . 9
3. RESEARCH OBJECTIVES . . . 9
   3.1 Technical Rationale . . . 10
4. METHODOLOGY . . . 10
   4.1 Overall Project Approach . . . 10
   4.2 Part 1: Bridge Plan Formatting Compliance Checking . . . 11
      4.2.1 System Overview . . . 11
      4.2.2 PDF Processing and Title Block Region Extraction . . . 11
      4.2.3 Visual Element Detection Using YOLOv5 . . . 13
      4.2.4 Optical Character Recognition (Split-OCR Approach) . . . 13
      4.2.5 Structured Field Extraction . . . 13
      4.2.6 OCR Error Correction . . . 13
      4.2.7 Compliance Rule Evaluation . . . 13
      4.2.8 Compliance Report . . . 14
   4.3 Part 2: Cross-Document Pay-Item Consistency Checking . . . 15
      4.3.1 Scope Confirmation and Data Objects . . . 15
      4.3.2 End-to-End Comparative Checking Workflow . . . 17
      4.3.3 Bridge Plan Item Extraction . . . 18
      4.3.4 Cost Estimate Item Extraction . . . 20
      4.3.5 Hybrid Matching and Consistency Checking . . . 21
      4.3.6 Data-Driven Algorithm Refinement . . . 23
   4.4 Web Application Integration and Deployment . . . 24
5. IMPLEMENTATION CONSIDERATIONS . . . 28
   5.1 Review-Assistance Use . . . 28
   5.2 Process and Data Integration . . . 28
   5.3 Algorithm Maintenance . . . 28
6. SUMMARY AND RECOMMENDATIONS . . . 28
   6.1 Summary and Findings . . . 28
   6.2 Recommendations and Implementation . . . 28
   6.3 Expected Benefits and Future Work . . . 29
REFERENCES . . . 29

LIST OF TABLES

Table 4.1	Validation Components Implemented.	11
Table 4.2	Included Checks and Requirement Descriptions.	11
Table 4.3	Title Block Crop Parameters Used During PDF Processing.	13
Table 4.4	YOLOv5 Model Configuration for Seal and Signature Detection.	15
Table 4.5	Compliance Field Extraction Methods.	15
Table 4.6	Summary of the Nine Compliance Checks Applied by the Rule Engine.	16
Table 4.7	Scope Decisions for Part 2 Comparative-Checking Module.	18
Table 4.8	Main Algorithm Modules in the Comparative Checking Workflow.	19
Table 4.9	Example Part 2 Matching and Checking Records.	22
Table 4.10	Development and Test Data Used for Part 2 Methodology Development.	23
Table 4.11	Development-Set Performance Changes for Bridge Plan Item Extraction.	24
Table 4.12	Development-Set Performance of Cost Estimate Item Extraction.	24
Table 4.13	Development-Set Performance Changes for Hybrid Matching and Consistency Checking.	24
Table 4.14	Examples of Data-Driven Refinements by Algorithm Component.	24
Table 4.15	Main Output Artifacts Preserved by the Workflow.	27

LIST OF FIGURES

Figure 4.1 Framework of Bridge Data Virtual Validator. 10

Figure 4.2 Overview of the Six-Stage Drawing Compliance Checker Pipeline. 11

Figure 4.3 Overview of the Drawing Compliance Checker Architecture. 12

Figure 4.4 Example Title Block Crop Extracted From a Full INDOT Engineering Drawing Page. 12

Figure 4.5 Custom Image Dataset Used for Training the YOLOv5 Detector and Labeling Process. 14

Figure 4.6 Example YOLOv5 Detection Results on INDOT Title Block Images. 15

Figure 4.7 Split-OCR Strategy Applied to INDOT Title Block Images. 15

Figure 4.8 Sample Compliance Report Output for an INDOT Plan Set. 16

Figure 4.9 Example Bridge Plan Summary Table Used as the Part 2 Checking Scope. 17

Figure 4.10 Example Pay-Item Information Located Outside the Bridge Plan Summary Table. 17

Figure 4.11 Example Cost Estimate Item Record Used for Cost-Side Pay-Item Extraction (With Sensitive Information Redacted). 18

Figure 4.12 Example Quantity Calculation Document Considered During Scope Confirmation. 19

Figure 4.13 Example Sheet List Used to Locate Summary Tables in a Bridge Plan. 20

Figure 4.14 Bottom-Region OCR Crop Used to Identify a Target Summary Page and Retain Designation Information. 20

Figure 4.15 Example Detected Bridge Plan Table Boundaries and Cells. 20

Figure 4.16 Organized OCR Text Sequence and Corresponding Bridge Plan Item Tuple. 21

Figure 4.17 Example of Parsing a Cost Estimate Item. 22

Figure 4.18 Designation-Based Pairing When One Bridge Plan is Associated With Multiple Cost Estimate Files. 22

Figure 4.19 Shared Review Spreadsheet Used to Summarize Algorithm Results and Review Comments. 23

Figure 4.20 Data-Driven Algorithm Development and Purdue/INDOT Review Loop. 23

Figure 4.21 Application Interface. 25

Figure 4.22 Bridge Plan Formatting Compliance Checker Interface. 25

Figure 4.23 PDF Upload Step for Compliance Checking. 25

Figure 4.24 PDF Size Validation Before Compliance Checking. 25

Figure 4.25 Running the Compliance Checking Process. 26

Figure 4.26 Compliance Checking Report Output. 26

Figure 4.27 Comparative Checking Upload Interface. 26

Figure 4.28 Interface of Comparative Checking Intermediate Outputs. 26

Figure 4.29 Example of Difference Table. 27

Figure 4.30 Interface of Downloading Comparative Checking Output Files. 27

Figure 4.31 Web Application Workflow for Project-Level Validation and Reviewer Interaction. 27

1. INTRODUCTION

Bridge projects are delivered through a collection of contract documents that describe design intent, quantities, payment items, project administration, and downstream asset information. For Indiana Department of Transportation (INDOT) bridge projects, Final Tracings Submittals include bridge plans, cost estimates, quantity-related information, Bridge Rating (BrR)–related information, and other supporting documents. These deliverables collectively serve as contractual records for letting and construction, and they also contain data that may be reused by bridge design, inspection, asset management, maintenance, and other downstream business processes (INDOT, 2024).

Bridge data validation has become more important as transportation agencies move toward digital delivery and more structured bridge information management. At the national level, bridge data validation emphasizes whether submitted bridge data satisfy data type, valid value, format, uniqueness, and presence requirements before being accepted into bridge inventory workflows (Federal Highway Administration [FHWA], 2025). At the state level, bridge quality control processes also recognize the need to check bridge contract drawings, design computations, and estimates before delivery decisions are made (New York State Department of Transportation [NYSDOT], 2021). These practices indicate that bridge deliverable quality assurance (QA) should address both document completeness and consistency among related data elements.

Recent research has demonstrated that automated methods can extract bridge information from drawings, tables, text blocks, and models (Akanbi & Zhang, 2022; Li & Zhang, 2025; Soy & Zhang, 2025). The National Cooperative Highway Research Program (NCHRP) Innovations Deserving Exploratory Analysis (IDEA) Project 230 applied machine learning and artificial intelligence to extract data and features from bridge plan sets (Shafei, 2024). Joint Transportation Research Program (JTRP) project SPR-4622 further demonstrated that OCR, natural language processing, machine learning, computer vision, and rule-based approaches can extract selected bridge information from two-dimensional (2D) bridge plans to support Specification for the National Bridge Inventory (SNBI)-related data collection (Zhang et al., 2026). These efforts provide an important technical foundation for the current JTRP project (SPR-4927), but their main focus was information extraction or model reconstruction rather than automated validation across multiple Final Tracings deliverables.

The current project extends this technical foundation from extracting individual bridge data elements to validating related information across project deliverables. The project investigates how existing Portable Document Format (PDF)- and model-based deliverables can be treated as document-derived data sources, how selected data objects can be extracted and structured, and how potential noncompliance items or cross-document discrepancies can be reported with traceable evidence. The initial implementation focuses on two validation functions: bridge plan format compliance checking and cross-document bridge data comparative checking.

2. PROBLEM STATEMENT

In the current Final Tracings workflow, bridge data are distributed across multiple documents and systems that are often managed independently. Although many deliverables are submitted electronically, the information contained in the files is not always reviewed in detail or structured for automated validation. As a result, related data elements may be repeated in different locations without a systematic mechanism to verify whether they are complete, valid, and consistent.

This creates a practical gap for INDOT. Existing federal and state validation practices, bridge plan information extraction tools, and building information modeling (BIM)-related workflows show that bridge data can be checked or extracted, but there is no INDOT-specific virtual validator for Final Tracings Submittals. In particular, a useful validation tool must operate on the documents that are currently submitted, identify which data objects can be checked, compare related information across deliverables, and produce traceable reports that support engineering judgment rather than automatic approval.

The problem addressed by this project is, therefore, how to convert selected information from existing bridge project deliverables into reliable validation records and use those records to support automated QA. This problem leads to the following research questions:

- What bridge data items in INDOT Final Tracings Submittals can or should be checked across bridge plans, cost estimates, and other related deliverables?
- How can optical character recognition (OCR), computer vision, natural language processing, machine learning, and rule-based reasoning be combined to extract and validate selected data items from existing documents?
- How can automated checking results be reported with enough source evidence and traceability to support human-in-the-loop engineering review?

3. RESEARCH OBJECTIVES

INDOT has recognized the need to refine the Final Tracings Submittal process for bridge projects because the submitted documents contain bridge data that support multiple downstream business functions, yet the contents of these documents are not always formally checked in a structured and automated manner. Building on SPR-4622 (Zhang et al., 2026) work on extracting bridge information from PDF plans and on the goal of cross-validating bridge data among Final Tracings documents, this project focuses on developing a Bridge Data Virtual Validator for INDOT bridge deliverable QA.

The overall objective of the current project is to develop and evaluate an automated validation tool that can extract selected data from existing bridge project deliverables, compare related information across documents, and generate traceable reports for engineering review. The specific objectives are:

- Review INDOT Final Tracings Submittals and related validation practices to identify bridge data items that can or should be checked across bridge plans, cost estimates, and other related deliverables.

- Develop a project-level validation framework and web application prototype that organizes bridge deliverable QA into bridge plan format compliance checking and bridge data comparative checking.
- Develop a bridge plan format compliance checking module that reads plan information and checks required control information, including file numbers, signatures, stamps, dates, and page numbering.
- Develop a cross-document comparative checking module that extracts, structures, matches, and compares pay-item information between bridge plans and cost estimates.
- Test the developed modules and refine the algorithms based on observed errors, and document test-set performance where gold standards are available.
- Provide traceable report outputs and implementation recommendations to support human-in-the-loop review, support future expansion to additional data objects, and downstream uses such as Indiana Total Assets Management Systems (iTAMS) data entry or bridge birth certificate workflows.

3.1 Technical Rationale

The technical rationale of this project is that many bridge QA checks do not require a complete replacement of current deliverable formats. Instead, selected information can be extracted from existing PDFs and model-related files, converted into structured validation records, and compared using transparent rules. This near-term approach is appropriate for INDOT because Final Tracings documents are already submitted electronically, while much of the information needed for review remains embedded in document layouts rather than stored in a unified database.

The proposed approach builds on prior bridge data extraction and digitalization efforts. The NCHRP IDEA Project 230 from Shafei (2024) demonstrated the use of machine learning and artificial intelligence to extract data and features from bridge plan sets, and SPR-4622 from Zhang et al. (2026) demonstrated OCR-, NLP-, computer vision-, and rule-based extraction of selected bridge plan information for SNBI-related use. These efforts support the feasibility of extracting document-embedded bridge data. The current project extends the application from extraction alone to validation and cross checking among project deliverables.

For bridge plan format compliance checking, the target evidence is largely visual and spatial. File numbers, page numbers, dates, signatures, and stamps appear in page-bottom tables or title-block regions rather than in uniform database fields. OCR systems provide a solid basis for recognizing text in document images (Du et al., 2020), while real-time object detection methods such as You Only Look Once (YOLO) support detection of visual elements and document regions (Wang et al., 2022). Rule-based reasoning is then appropriate for checking whether required evidence is present, whether file numbers are consistent, and whether page numbers are continuous. This combination supports explainable compliance checking because each noncompliance item can be linked back to a source page region.

For cross-document comparative checking, the target information is semistructured rather than purely visual. Bridge plan pay items are embedded in quantity tables, whereas cost estimate pay items are organized as itemized records. Therefore, OCR and layout processing are needed to recover item names, units, quantities, and source metadata from plans, while text segmentation and parsing are needed for cost estimate records. After extraction, exact string matching is insufficient because bridge plans and cost estimates may use different abbreviations, wording, and/or levels of detail. A hybrid matching strategy that combines normalization, lexical similarity, semantic correction, and deterministic unit-and-quantity checking provides a transparent way to detect discrepancies while preserving review evidence.

The project follows a reviewer-centered validation philosophy. Automated checking is intended to prescreen deliverables and narrow the scope of manual review, not to replace engineering judgment. For discrepancy detection, recall is especially important because a missed inconsistency may remain in the deliverable package. Precision remains important, but false alarms can be resolved if the report preserves source evidence, matching explanations, and reviewer status fields. This rationale supports the use of data-driven algorithm refinement based on INDOT projects and the design of a web application that allows engineers to inspect, confirm, dismiss, or comment on reported issues.

4. METHODOLOGY

4.1 Overall Project Approach

The Bridge Data Virtual Validator was developed as a project-level QA workflow for validating INDOT bridge project deliverables. As shown in Figure 4.1, the methodology is organized around two complementary validation functions, namely compliance checking (Part 1) and comparative checking (Part 2), respectively. Part 1 checks whether required letting processing information is present in bridge plans and internally consistent. Part 2 checks whether bridge data are consistent across multiple deliverables. In both parts, the validator converts document evidence into structured records before applying rule-based or hybrid comparison logic.

The same design principle was used across the project: automated checking should produce traceable review evidence

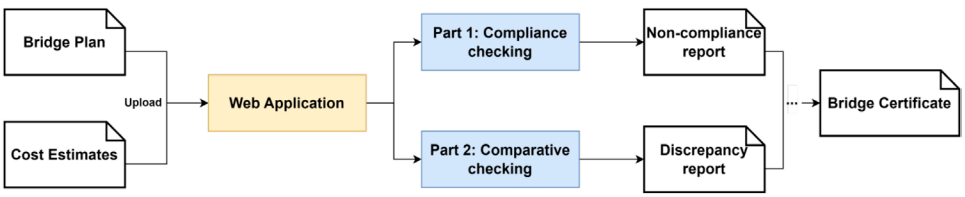


Figure 4.1 Framework of Bridge Data Virtual Validator.

rather than mere pass/fail decisions. Each generated record, therefore, keeps the source file, page or text block, extracted fields, matching result, and issue note needed for engineering review. Table 4.1 shows the object of output of each component.

TABLE 4.1
Validation Components Implemented.

Component	Primary Data Object	Current Output
Part 1: Bridge plan format compliance checking	Page-bottom title-block/control information in bridge plans	Noncompliance report for missing signatures, dates, stamps, inconsistent file numbers, and page-number continuity
Part 2: Cross-document pay-item consistency checking	Pay items in bridge plan summary tables and cost estimate item records	Consistency report with consistent, inconsistent, not-found, and additional cost-estimate records
Web application integration	Project files, intermediate outputs, and review reports	Traceable review package for human-in-the-loop QA

TABLE 4.2
Included Checks and Requirement Descriptions.

No.	Check	Requirements
1	Check PDF page size	Must be landscape orientation, 36″ × 24″ or 34″ × 22″
2	Bridge file number consistency	Must have a value and be consistent across all pages. None/Null/N/A is acceptable as long as it is consistent.
3	Designation number consistency	Must have a designation number and be consistent across all pages.
4	Contract number consistency	Must have a contract number and be consistent across all pages.
5	Total sheet count consistency	Must have a total sheet number and be the same across all pages.
6	Sheet number sequence	Must have a sheet number and must be sequential.
7	Date	Must be present on all pages except cross-section pages.
8	Signature presence	Must be present on all pages except cross-section pages.
9	Indiana Seal presence	Must be present on all pages except cross-section pages.

4.2 Part 1: Bridge Plan Formatting Compliance Checking

4.2.1 System Overview

The Drawing Compliance Checker is an automated pipeline developed to check INDOT bridge engineering drawing PDFs against a set of compliance requirements shown in Table 4.2. Engineering plan sets submitted for contract services consist of multipage PDF documents, where each page contains a standardized INDOT title block. This title block carries critical project information including the designation number, contract number, bridge file number, sheet number, total sheet count, date, professional engineer (PE) signature, and the Indiana state seal. The system automatically extracts and verifies all these fields across every page of a submitted plan set.

The pipeline is structured as a six-stage sequential process, as illustrated in Figure 4.2. Each stage produces an intermediate output that serves as input to the next stage. The six stages are:

1. PDF processing and title block region extraction,
2. Visual element detection using YOLOv5,
3. OCR using a split-image approach,
4. Structured field extraction from OCR results,
5. OCR error correction and adjustment, and
6. Compliance rule evaluation and report generation.

The entire pipeline is wrapped in a Streamlit-based web application. The overall system architecture is shown in Figure 4.3.

4.2.2 PDF Processing and Title Block Region Extraction

Each page of the submitted PDF is rendered as a rasterized image using PyMuPDF (fitz) at a resolution of 300 dots per inch (DPI). At this resolution, the title block region occupies approximately 5,600 × 1,050 px on a standard INDOT engineering drawing. Rather than processing the full page, which contains richer drawing content, some of which could be irrelevant to compliance, a fixed-region crop is applied to isolate the region of interest block (shown in Figure 4.4). This approach followed the A3 INDOT drawing standards to prescribe that the title block always occupies the bottom-right corner of the sheet for a landscape layout.

The crop region is defined by computing offsets from the bottom-right edge of the rendered image converted from the

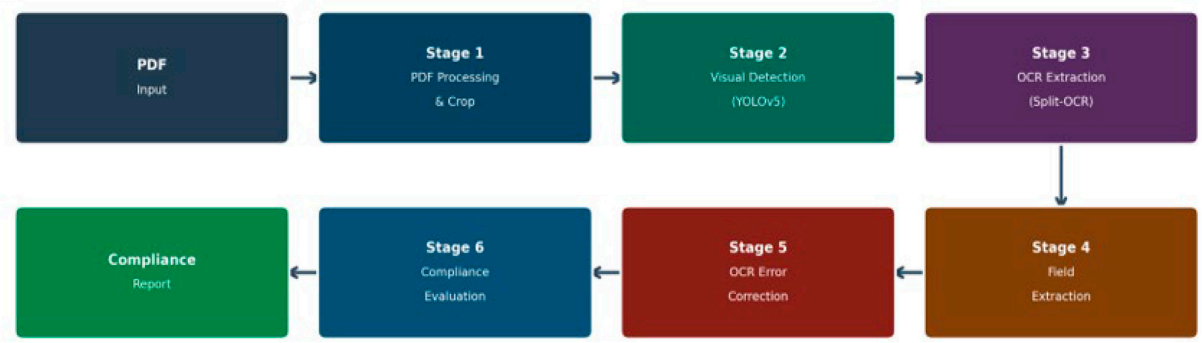


Figure 4.2 Overview of the Six-Stage Drawing Compliance Checker Pipeline.

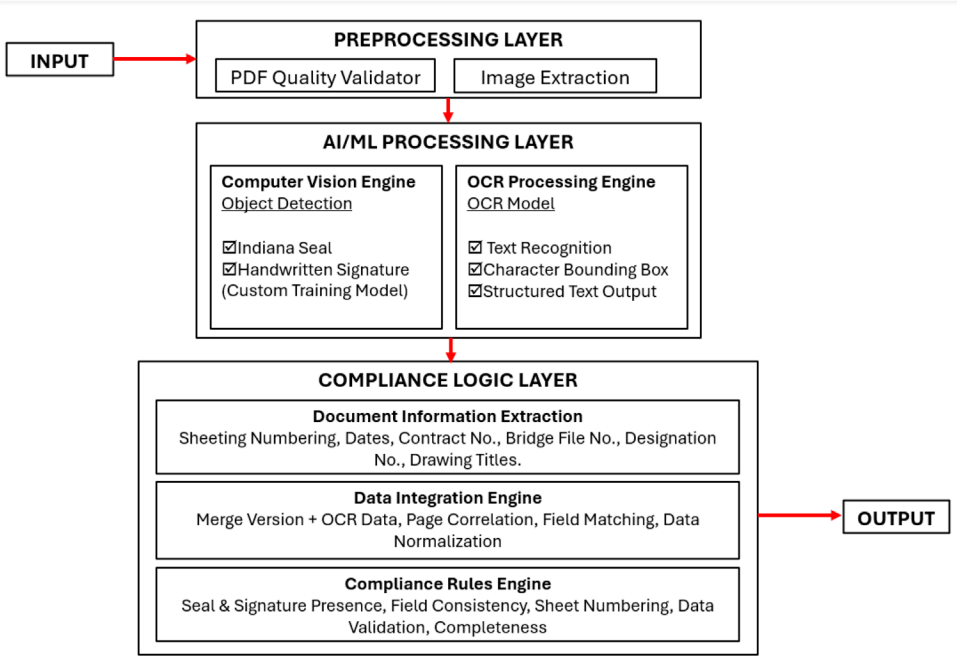


Figure 4.3 Overview of the Drawing Compliance Checker Architecture.

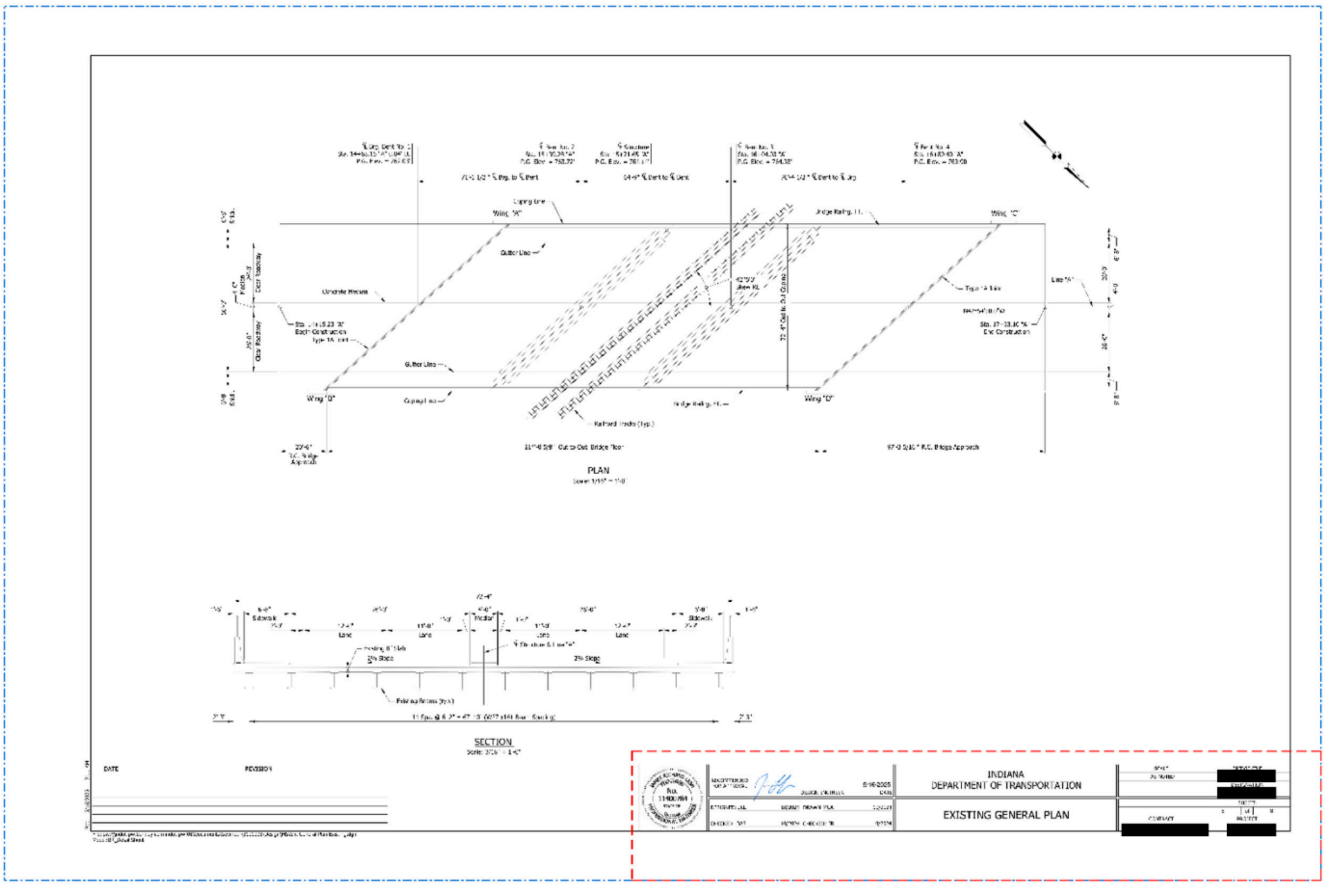


Figure 4.4 Example Title Block Crop Extracted From a Full INDOT Engineering Drawing Page.

TABLE 4.3
Title Block Crop Parameters Used During PDF Processing.

Parameter	Value	Description
DPI	300	Rendering resolution (balances image quality and processing speed)
CROP_W	5,600 px	Width of title block crop region at 300 DPI
CROP_H	1,050 px	Height of title block crop region at 300 DPI
MARGIN	50 px	Inset from page edge to avoid border artifacts

drawing page. A 50-px margin is applied from the page boundary to avoid capturing border artifacts. The resulting crop dimensions are 5,600 px (width) by 1,050 px (height). The cropped title block images are saved as individual PNG files, one per PDF page, and passed to subsequent pipeline stages. Table 4.3 provides the preprocessing and arrangement parameters of those image processing.

4.2.3 Visual Element Detection Using YOLOv5

Two visual compliance elements (the Indiana state seal and the professional engineer's handwritten signature) cannot be reliably detected through text recognition, as they are graphical rather than textual information. Therefore, You Only Look Once (YOLO), which is a powerful deep-learning-based real-time visual detection model, was exploited to support the seal and signature detection. A custom YOLOv5 object detection model was trained to detect these two classes within each cropped title block image. Figure 4.5 shows the dataset and labeling process of these two classes (Indiana seal and Signature).

The model was trained on a dataset of labeled INDOT title-block images containing annotated instances of the Indiana state seal and PE signatures. Training was conducted using the YOLOv5 architecture with transfer learning from pretrained weights. The trained model weights are stored as a single file (best.pt) and loaded on first use, remaining in memory for the duration of a pipeline run.

Prior to inference, each title-block image is upscaled by a factor of 2 using bicubic interpolation. This brings an effective resolution from 300 DPI to an equivalent of 600 DPI, which improves detection sensitivity for small seal impressions and faint handwritten signatures that may be difficult to resolve at the original resolution. The confidence threshold for detection is set to 0.1 to minimize false negatives. A page is marked as having a detected element if at least one bounding box for that class exceeds the confidence threshold. Table 4.4 shows the rationale behind the selected parameters and processing in these two classes of detection. Figure 4.6 shows an example of the detection of Indiana seal and signature on a cropped section of the drawing page.

4.2.4 Optical Character Recognition (Split-OCR Approach)

Textual information in the title block is extracted using a deep-learning-based OCR engine.

A key technical challenge arises from the dimensions of the title-block crop. When the full image is fed directly to the OCR engine, the engine internally downscales it to approximately 71% of its original resolution, which causes degradation of small or low-resolution handwritten text, particularly date fields and designation numbers that appear in narrow cells on the right-hand side of the title block. To address this, a Split-OCR strategy was developed. Each title block is divided into two overlapping crops—a left crop covering the project description area and a right crop covering the critical compliance fields—which are processed independently through the OCR engine. Detections from the overlapping zone are deduplicated using Intersection over Union (IoU) comparison, and the final output is a merged list of all unique text items with their bounding box coordinates and confidence scores. Figure 4.7 illustrates the Split-OCR approach.

4.2.5 Structured Field Extraction

Following OCR extraction, a set of spatial rule-based algorithms processes the text item list for each page to locate and extract specific compliance fields. Each field has a dedicated extractor that uses anchor-based spatial reasoning: a known label (e.g., DESIGNATION, CONTRACT, SHEETS) is first located in the OCR output, and candidate values are collected from the spatially adjacent region. Normalized x/y coordinates ensure comparisons are resolution independent. A vertical text filter discards rotated annotations that could interfere with spatial matching. Table 4.5 summarizes the extraction approach for each compliance field.

4.2.6 OCR Error Correction

Engineering drawings printed at standard resolution occasionally produce OCR misreadings in sheet number fields due to the visual similarity between the digits 6 and 9 when text is small or lightly inked. Because these digits are rotationally symmetric, an OCR engine viewing a '9' that is slightly misaligned may return '6,' and vice versa. An automated correction stage is applied after the merged comma-separated values (CSV) is produced to detect and correct these errors before compliance evaluation.

For the total sheet count, a three-signal check compares the PDF page count, the statistically expected total derived from the sheet number sequence, and the most commonly occurring value across all pages. Where two or more signals agree and the outlying value is a 6/9 flip of the expected total, the correction is applied automatically.

For individual sheet numbers, the expected value for each page is computed from the implied first sheet and the page position. Where the OCR value differs from the expected and the difference is consistent with a 6/9 digit flip, the corrected value is substituted. All corrections are logged with the page number, original value, corrected value, and reason, and are included in the final compliance report.

4.2.7 Compliance Rule Evaluation

The compliance rule engine applies nine checks organized into four categories: (1) a document format check verifying PDF page dimensions; (2) consistency checks verifying that the bridge file number, designation number, contract number, and total sheet count are identical across all pages; (3)

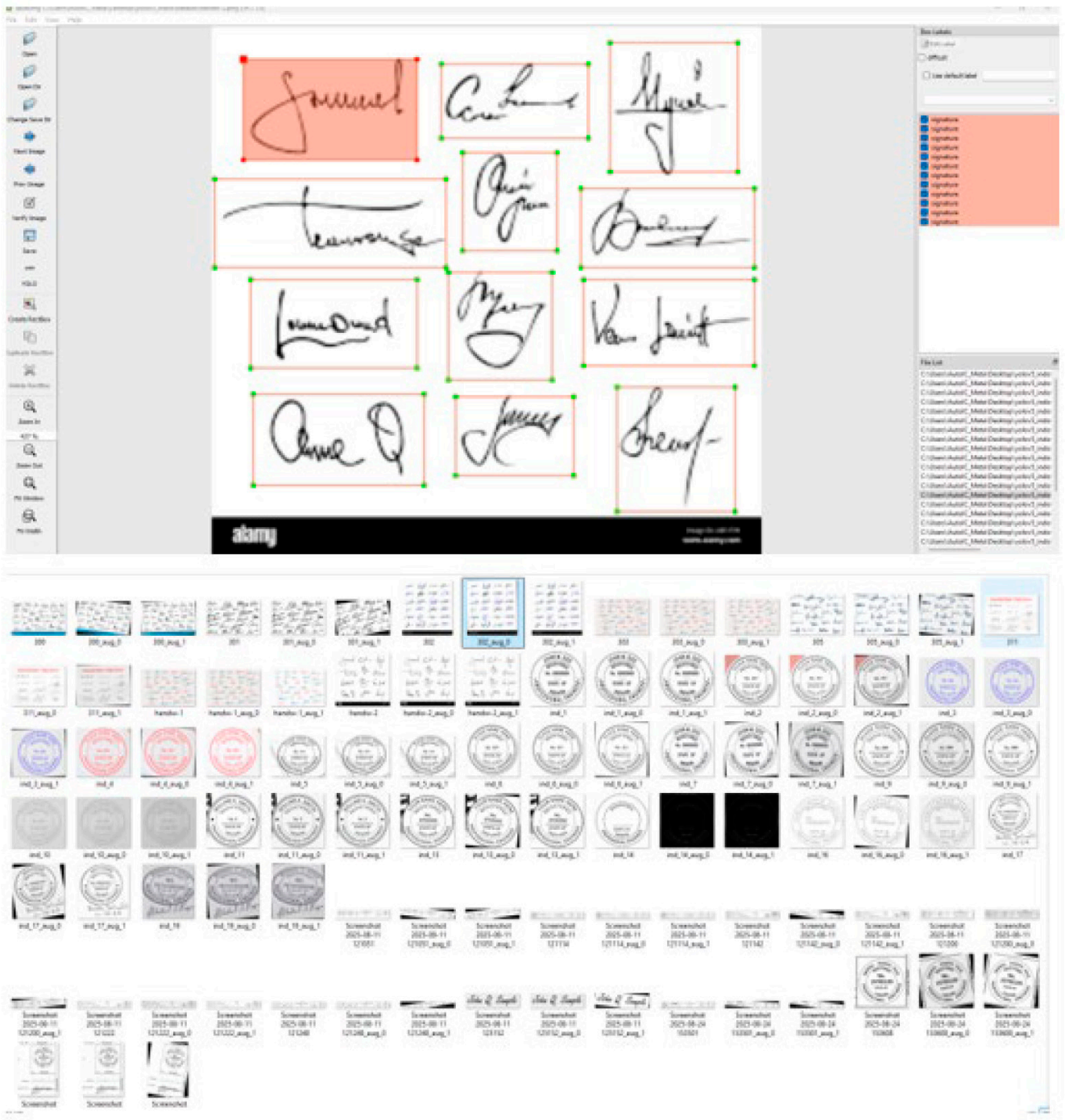


Figure 4.5 Custom Image Dataset Used for Training the YOLOv5 Detector and Labeling Process.

a sequential order check verifying sheet numbers increase without gaps; and (4) visual presence checks verifying that a PE signature, Indiana state seal, and date are present on all applicable pages. Cross-section pages are automatically identified and excluded from visual element checks. Each check produces a binary PASS or FAIL result with a supporting detail message identifying the specific pages or values responsible for any noncompliance (Table 4.6).

4.2.8 Compliance Report

The compliance report is generated as a plain text file and saved alongside the intermediate pipeline outputs. The report lists the result of every compliance check using pass (✓) or fail (✗) indicators, identifies specific page numbers for any failures, and includes a summary section enumerating all issues found. Any OCR corrections applied in Stage 5 are also listed

TABLE 4.4
YOLOv5 Model Configuration for Seal and Signature Detection.

Parameter	Value	Rationale
Detection classes	indiana_seal, signature	Two visual compliance elements
Confidence threshold	0.1	Low threshold minimizes false negatives
Input pre-processing	2× bicubic upscale	Improves resolution for small elements
Inference device	CPU	Deployment server has no GPU
Model weights	best.pt (custom trained)	Trained on INDOT title block images

in the report header with the original and corrected values, providing full traceability. The report filename follows the pattern: <document_name>_compliance_report.txt, shown in Figure 4.8.

4.3 Part 2: Cross-Document Pay-Item Consistency Checking

4.3.1 Scope Confirmation and Data Objects

First, the project team investigated the collected deliverables and reviewed which of them should be checked for cross-document bridge data consistency. The original candidate scope included bridge plans, cost estimates, and quantity

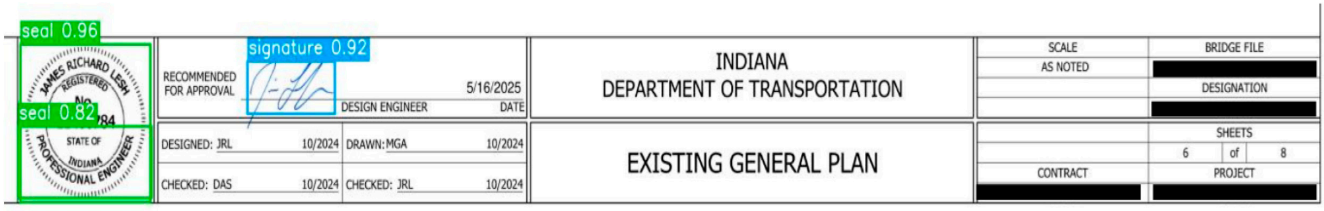


Figure 4.6 Example YOLOv5 Detection Results on INDOT Title Block Images.

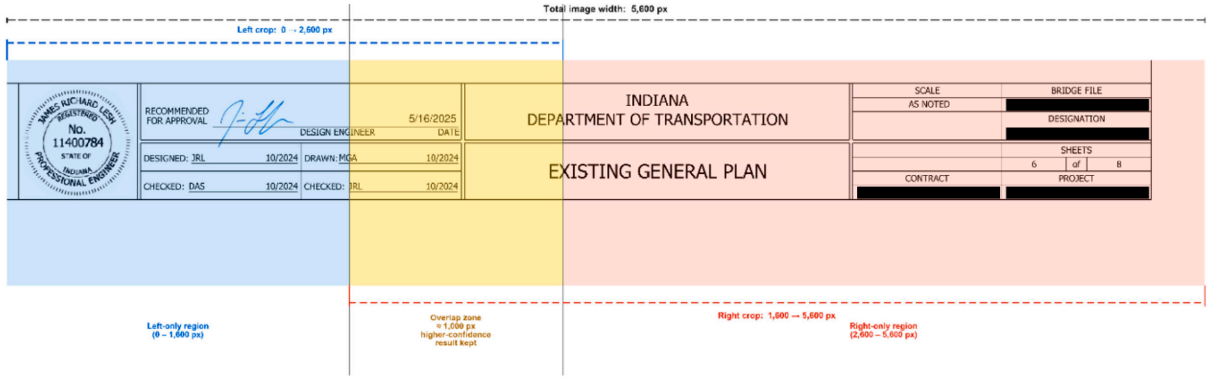


Figure 4.7 Split-OCR Strategy Applied to INDOT Title Block Images.

TABLE 4.5
Compliance Field Extraction Methods.

Field	Anchor Label	Search Region	Notes
Designation Number	DESIGNATION	200 px below label, ±500 px horizontal	6–8 digit number; N/A treated as fallback only
Contract Number	CONTRACT (exact match only)	Below label ±100 px horizontal, +200 px vertical	Pattern: [A-Z]-XXXX to XXXXXX (e.g., T-44569, B-40589)
Bridge File Number	BRIDGE FILE (or split BRIDGE + FILE tokens)	±300 px horizontal, +300 px vertical	Pattern: XXX-XX-XXXXX with optional letter suffix; also (N)X-XX-XXXXX
Date	DATE (also 'CTE','DTE')	500 px ABOVE label, 1000 px left, 500 px right	Date pattern; applies char corrections (F→7, I→1, O→0)
Current Sheet	SHEETS (x_normalized≥0.85)	700 px horizontal, 200 px vertical	Leftmost digit candidate across all matches
Total Sheets	SHEETS (x_normalized≥0.85)	700 px horizontal, 200 px vertical	Rightmost digit candidate across all matches
Sheet Title	DEPARTMENT OF TRANSPORTATION	Zone below label	All text in zone, joined by row
Signature (OCR)	RECOMMENDED	Right of label (Layout 1); Left of label (Layout 2)	PE name text — stored for reference only

TABLE 4.6
Summary of the Nine Compliance Checks Applied by the Rule Engine.

#	Compliance Check	Category	Pages Evaluated	Pass Condition
1	PDF Page Size	Document Format	All pages	Landscape orientation; 36″ × 24″ or 34″ × 22″
2	Bridge File Number	Consistency	All pages	Same value on all pages; None/N/A acceptable if consistent
3	Designation Number	Consistency	All pages	Same value on all pages; none empty
4	Contract Number	Consistency	All pages	Same value on all pages; none empty
5	Total Sheet Count	Consistency	All pages	Same value on all pages; none empty
6	Sheet Number Sequence	Sequential	All pages	Sheet number present; strictly sequential, no gaps or duplicates
7	Date	Visual	Non-cross-section pages	Date present and matches valid date format
8	Signature Presence	Visual	Non-cross-section pages	PE signature detected by YOLO
9	Indiana Seal Presence	Visual	Non-cross-section pages	Indiana state seal detected by YOLO

```
COMPLIANCE REPORT (v2 — CSV-based)
Document: [redacted]
Total pages: 95
Cross-section pages excluded from visual checks: 0
============================================================

❌ Designation Number: inconsistent
    '[redacted]035' (majority) — 41 pages
    '[redacted]037' — 30 pages: 24, 28, 29, 30, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92,
93, 94, 95
    '1800035/1800037' — 24 pages: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 27, 32

✅ Contract Number: R-[redacted] (95 pages — all consistent)

❌ Bridge File Number: inconsistent
    'N.A.' (majority) — 51 pages
    '(empty)' — 44 pages: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29,
30, 31, 32, 34, 35, 40, 41, 42, 47, 48, 53, 54, 59, 60, 65

✅ Total Sheets: 565 (95 pages — all consistent)

✅ Sheet Order: sheets 1–95 in correct order

✅ Date: all 95 checked pages pass  (0 cross-section pages excluded)

✅ Signature: all 95 checked pages pass  (0 cross-section pages excluded)

✅ Indiana Seal: all 95 checked pages pass  (0 cross-section pages excluded)

============================================================
SUMMARY
----------------------------------------
❌ NOT COMPLIANT — 2 issue(s):
  • Designation Number inconsistent
  • Bridge File Number inconsistent
```

Figure 4.8 Sample Compliance Report Output for an INDOT Plan Set.

calculations. During scope confirmation, quantity calculations were excluded from the implemented checking target because they are commonly used as an intermediate manual calculation artifact, as shown in Figure 4.12. In contrast, bridge plans present data that are related to construction directly, and cost estimates organize item records for payment, as shown in Figure 4.11. Therefore, the implemented workflow compares bridge plan data against cost estimate records with the same project designation number.

The bridge data considered in the workflow was also scoped to pay items listed in summary tables, such as bridge summary, road summary, summary tables, summary of quantities, and roadway summary sheets, as shown in Figure 4.9. Pay items that appear only as scattered callouts or notes in plan drawings, as shown in Figure 4.10, were not included in the current automated checking target because they require a different extraction strategy and are less consistently organized across plan sets. Table 4.7 summarizes the scope.

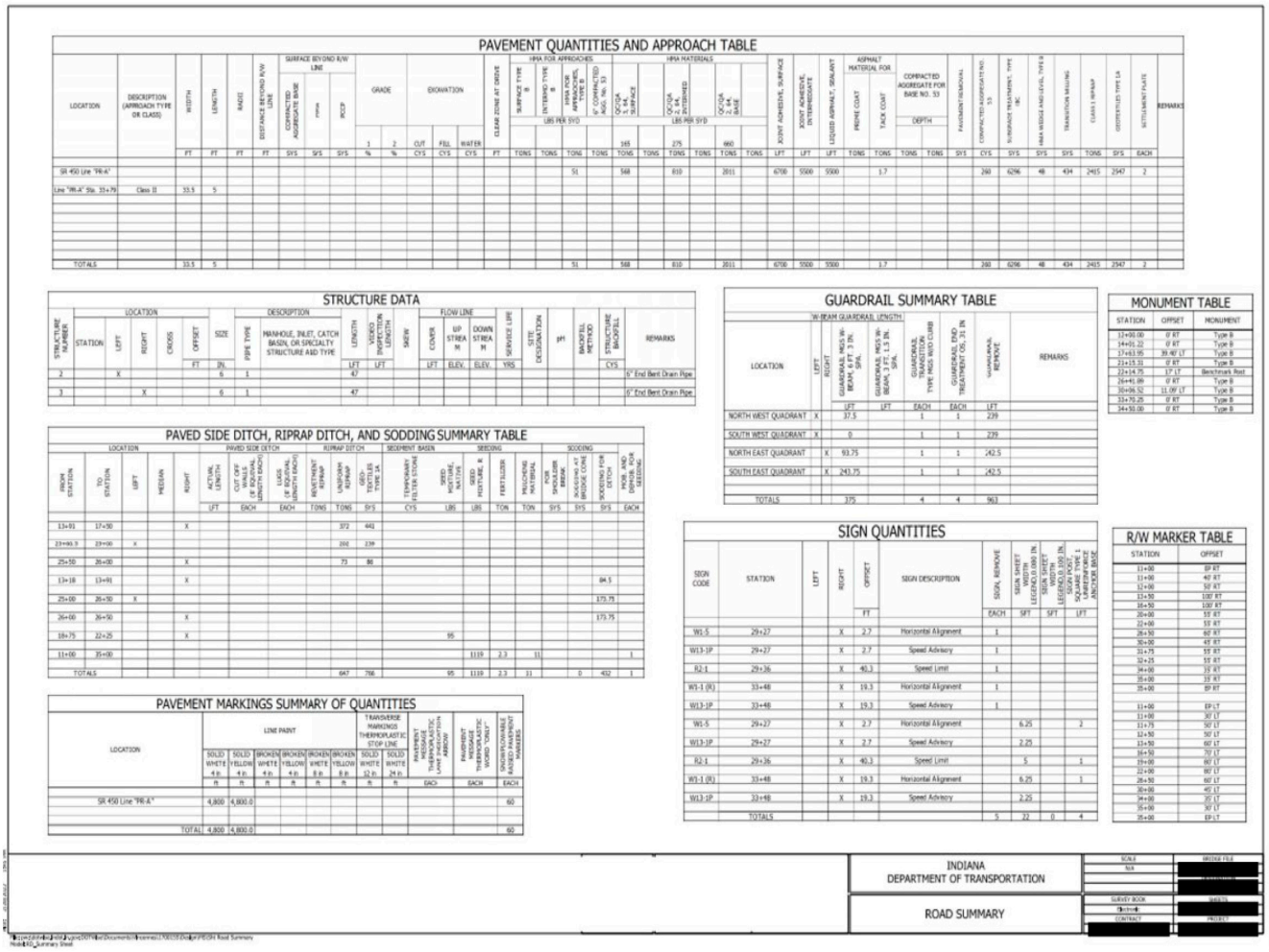


Figure 4.9 Example Bridge Plan Summary Table Used as the Part 2 Checking Scope.

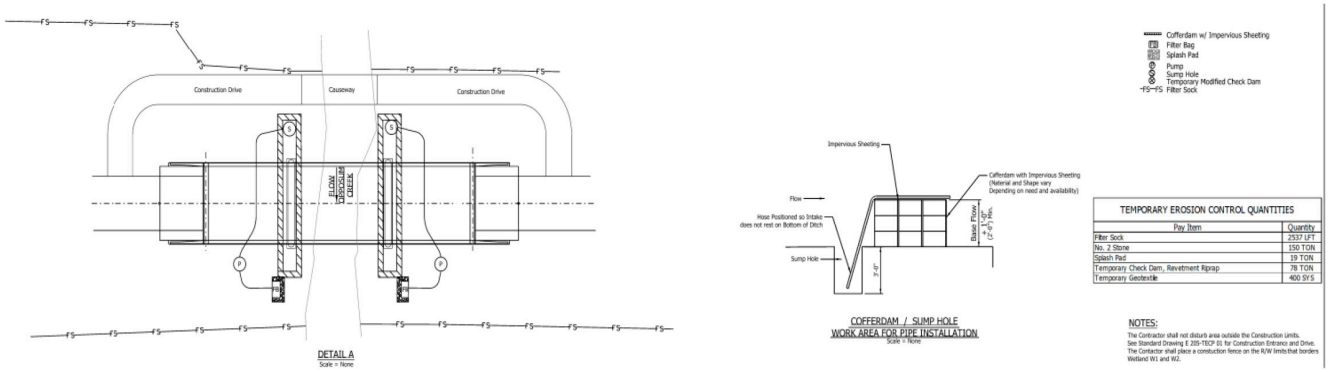


Figure 4.10 Example Pay-Item Information Located Outside the Bridge Plan Summary Table.

4.3.2 End-to-End Comparative Checking Workflow

After the scope was fixed, the comparative checking workflow was implemented as three connected algorithm modules: bridge plan item extraction, cost estimate item extraction, and hybrid matching. In the bridge plan branch, the algorithm first screens plan pages using bottom-region OCR and then performs image-based OCR on selected summary pages, followed by rule-based structuring to organize the extracted items. In the cost estimate branch, text-based OCR and regular-expression parsing are

TABLE 4.7
Scope Decisions for Part 2 Comparative-Checking Module.

Candidate Deliverable	Role in Project Delivery	Treatment in Comparative Checking
Bridge plan	Construction-related information including various bridge data	Included. Only pay items listed in target summary tables are extracted and checked.
Cost estimate	Payment records in a standard estimate format	Included. Treated as the reference item set for name matching and unit/quantity checking.
Quantity calculation	Manual or semi-manual intermediate quantity calculation	Not included in current automated checking. Retained as a future validation source.
Scattered bridge data and construction notes	Design details distributed across documents	Not included in the current workflow

End-to-End Comparative Checking Workflow

INDIANA DEPARTMENT OF TRANSPORTATION — DATE : 02/25/2022
CES JOB SUMMARY ESTIMATE — PAGE : 4

JOB NUMBER: — ESTIMATOR: — SPEC YEAR:
DESCRIPTION: — COUNTY: — WORK TYPE:
SEASON: — UNIT SYSTEM:

HIGHWAY TYPE: — CONTINGENCY:
DATE UPDATED: — LETTING DATE: — READY FOR CONTRACT DATE:
CONTRACT NO: — CHECKED BY: — DATE CHECKED:
LATITUDE: — PROJECT LENGTH: — PAVEMENT WIDTH:
LONGITUDE: — LANE MILES/KM: — PAVEMENT DEPTH:

LINE	CAT	ITEM	DESCRIPTION	QTY	UNIT		PRICE		AMOUNT	OBS	REC/STD/UNIQUE
0049	0001	628-11976	COMPUTER SYSTEM EQUIPMENT	2.0000	EACH	@	$ 887.69	=	$ 1,775.39	N	S
0050	0001	628-09403	FIELD OFFICE C	14.0000	MOS	@	$ 2,597.37	=	$ 36,363.28	N	S
0052	0001	701-06011	DYNAMIC PILE LOAD TEST	2.0000	EACH	@	$ 3,272.00	=	$ 6,544.00	N	S
0053	0001	701-09557	TEST PILE DYNAMIC PRODUCTION	174.0000	LFT	@	$ 50.00	=	$ 8,700.00	N	S
0054	0001	701-09559	TEST PILE DYNAMIC RESTRIKE	2.0000	EACH	@	$ 2,129.33	=	$ 4,258.66	N	S
0055	0001	701-09891	PILE SHOE HP 12 X 84	28.0000	EACH	@	$ 140.27	=	$ 3,927.70	N	S
0056	0001	701-95782	PILE STEEL H HP 12 X 84	2,000.0000	LFT	@	$ 100.21	=	$ 200,425.54	N	S
0057	0001	702-12076	GRATES BASINS AND FITTINGS CAST IRON	3.0000	EACH	@	$ 1,473.84	=	$ 4,421.53	N	S
0058	0001	702-51005	CONCRETE A SUBSTRUCTURE	137.4000	CYS	@	$ 1,145.49	=	$ 157,390.33	N	S
0059	0001	703-06028	REINFORCING BARS	22,130.0000	LBS	@	$ 1.02	=	$ 22,722.20	N	S 703-R-724
0060	0001	703-06029	REINFORCING BARS EPOXY COATED	57,214.0000	LBS	@	$ 1.19	=	$ 68,268.32	N	S 703-R-724
0061	0001	703-97936	THREADED TIE BAR ASSEMBLY EPOXY COATED	36.0000	EACH	@	$ 36.19	=	$ 1,302.97	N	S 703-R-724
0062	0001	704-51002	CONCRETE C SUPERSTRUCTURE	174.7000	CYS	@	$ 939.01	=	$ 164,045.49	N	S
0063	0001	706-11600	RAILING CONCRETE FC	24.1000	CYS	@	$ 937.52	=	$ 22,594.39	N	S
0064	0001	706-11620	CONCRETE BRIDGE RAILING TRANSITION TFC	4.0000	EACH	@	$ 1,955.28	=	$ 7,821.12	N	S
0065	0001	707-05984	STRUCTURAL MBR CONC I-BEAM 45 X 16 IN	498.0000	LFT	@	$ 218.56	=	$ 108,845.89	N	S
0066	0001	709-51821	SURFACE SEAL	1.0000	LS	@	$ 2,000.00	=	$ 2,000.00	N	S BRIDGE NO IN SUPP DESC
0067	0001	715-05407		94.0000	LFT	@	$ 36.08	=	$ 3,391.83	N	S

Figure 4.11 Example Cost Estimate Item Record Used for Cost-Side Pay-Item Extraction (With Sensitive Information Redacted).

performed. The two structured item sets are then aligned through a hybrid matching strategy and checked for consistency item by item. The final report presents results with traceable evidence from source documents. Table 4.8 summarizes the main modules.

4.3.3 Bridge Plan Item Extraction

Bridge plan item extraction converts selected bridge plan summary pages into records of the form: bridge item = (item name, unit, quantity, source evidence). The source evidence includes the original file, page number, table identifier, and OCR-derived text/cell information. The designation number identified during page screening is retained with source evidence so that bridge plan items can later be aligned with cost estimate files.

The first step is OCR-based target-page screening. During the investigation, it was identified that pay items are typically summarized in a few pages that contain summary tables of pay item information, as shown in Figure 4.13. Running full OCR on each page is tedious and unnecessary. Therefore, the algorithm first selects and slices the target page.

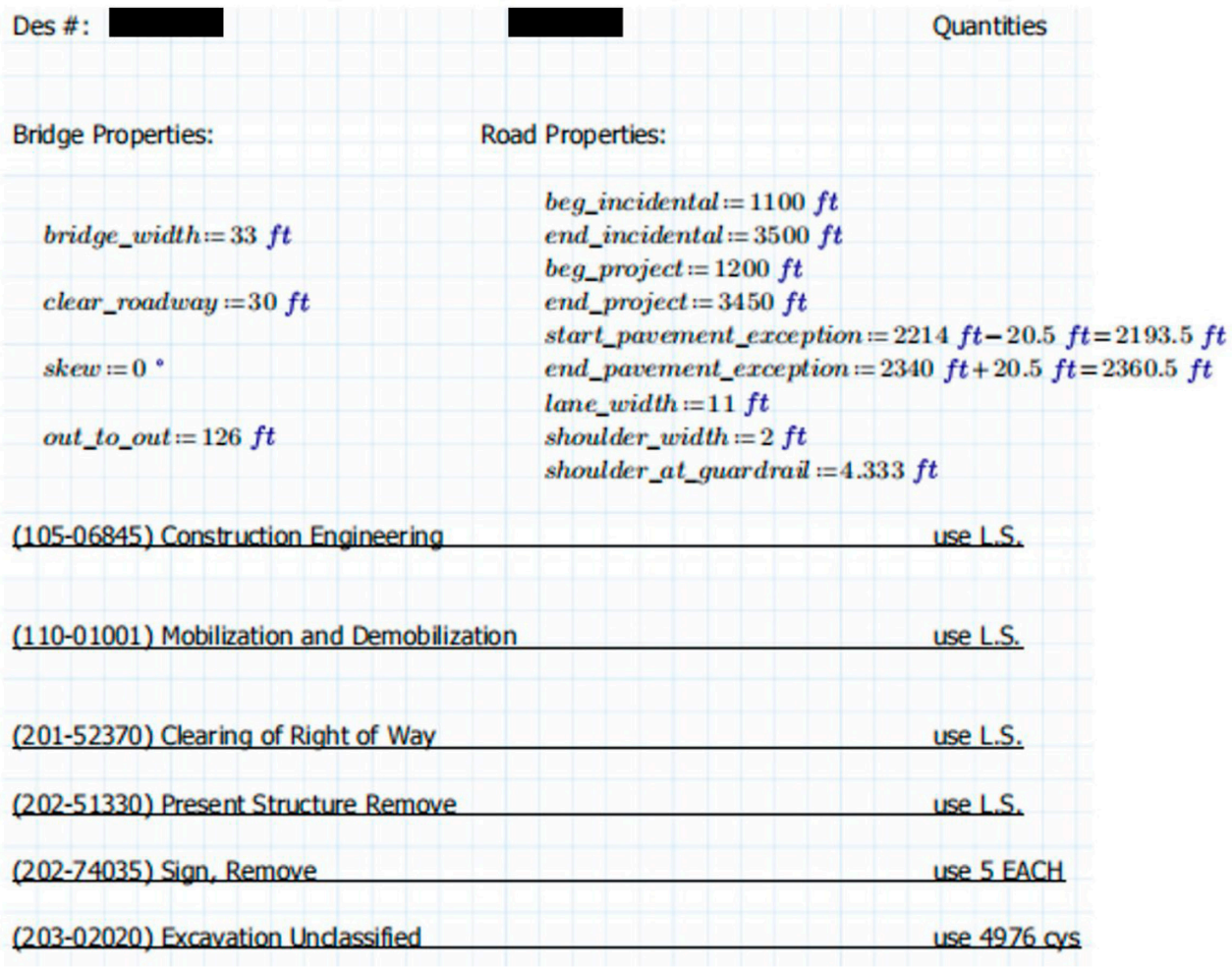


Figure 4.12 Example Quantity Calculation Document Considered During Scope Confirmation.

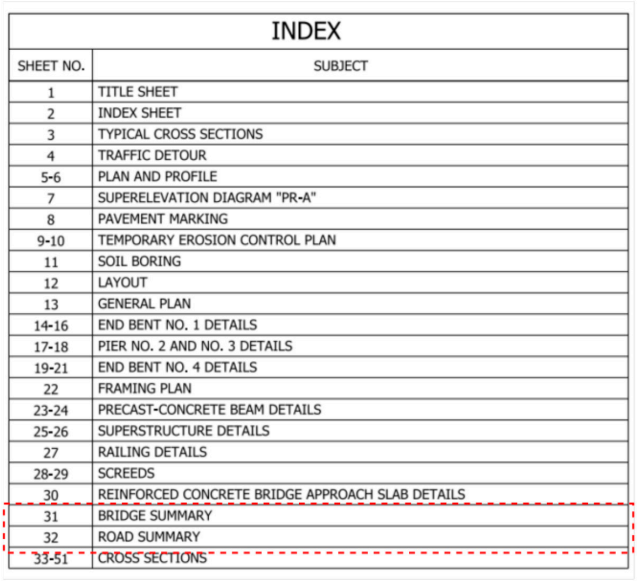
TABLE 4.8
Main Algorithm Modules in the Comparative Checking Workflow.

Module	Input	Output
Bridge plan item extraction	Bridge plan PDF	Structured bridge item records
Cost estimate item extraction	Cost estimate PDF	Structured cost estimate item records
Hybrid matching and checking	Bridge and cost item records grouped by designation number	Consistency checking report
Report generation	Checking results and source metadata	Reviewable discrepancy package

Bridge Plan Item Extraction

For each bridge plan page, the algorithm renders the bottom-right title-block region rather than the full page, as shown in Figure 4.14. The crop is preprocessed using grayscale conversion, contrast enhancement, denoising, sharpening, and Sauvola thresholding before OCR. The normalized OCR text is compared with a predefined target title set including titles such as bridge summary, road summary, summary tables, summary of quantities, and roadway summary. Pages that contain a target title are sliced into single-page PDFs for full table extraction.

For each selected summary page, the algorithm renders the page at high resolution and detects table and cell boundaries using OpenCV-based line extraction. The algorithm uses horizontal and vertical morphological kernels to recover table skeletons, connected components to retain table-line candidates, and contour-based bounding boxes to define table cells. An example of detected cells is shown in Figure 4.15, and the process of structuring a pay item is shown in Figure 4.16.



INDEX

SHEET NO.	SUBJECT
1	TITLE SHEET
2	INDEX SHEET
3	TYPICAL CROSS SECTIONS
4	TRAFFIC DETOUR
5-6	PLAN AND PROFILE
7	SUPERELEVATION DIAGRAM "PR-A"
8	PAVEMENT MARKING
9-10	TEMPORARY EROSION CONTROL PLAN
11	SOIL BORING
12	LAYOUT
13	GENERAL PLAN
14-16	END BENT NO. 1 DETAILS
17-18	PIER NO. 2 AND NO. 3 DETAILS
19-21	END BENT NO. 4 DETAILS
22	FRAMING PLAN
23-24	PRECAST-CONCRETE BEAM DETAILS
25-26	SUPERSTRUCTURE DETAILS
27	RAILING DETAILS
28-29	SCREEDS
30	REINFORCED CONCRETE BRIDGE APPROACH SLAB DETAILS
31	BRIDGE SUMMARY
32	ROAD SUMMARY
33-51	CROSS SECTIONS

Figure 4.13 Example Sheet List Used to Locate Summary Tables in a Bridge Plan.

4.3.4 Cost Estimate Item Extraction

Compared with bridge plans, cost estimate PDFs are text-based and organized by lines and text blocks instead of tables. The algorithm converts each cost estimate PDF page to an image, crops the left portion of the page where the item block is located, preprocesses the image using contrast enhancement, denoising, sharpening, and Sauvola thresholding, and applies OCR with a line-oriented page segmentation mode. The resulting lines are cleaned and stitched across pages before item records are parsed. Figure 4.17 shows an example of parsing a cost estimate item.

Cost estimate items are normalized using the same item-name and unit-normalization functions applied to bridge plan items. When a bridge plan designation corresponds to more than one cost estimate file, the cost items are combined before matching. Records with the same normalized name and normalized unit are summed, while records with the same normalized name but different units are preserved as separate candidates so that possible unit conflicts are not hidden. Figure 4.18 shows an example of combining a cost estimate item.

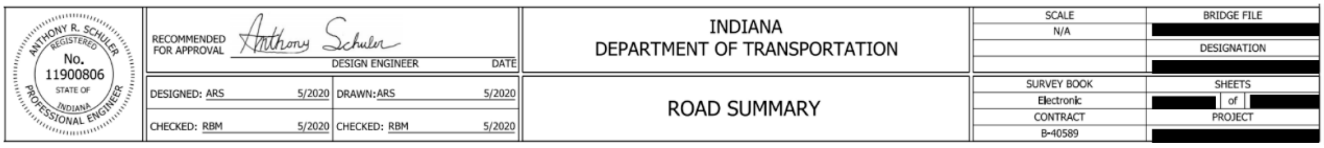
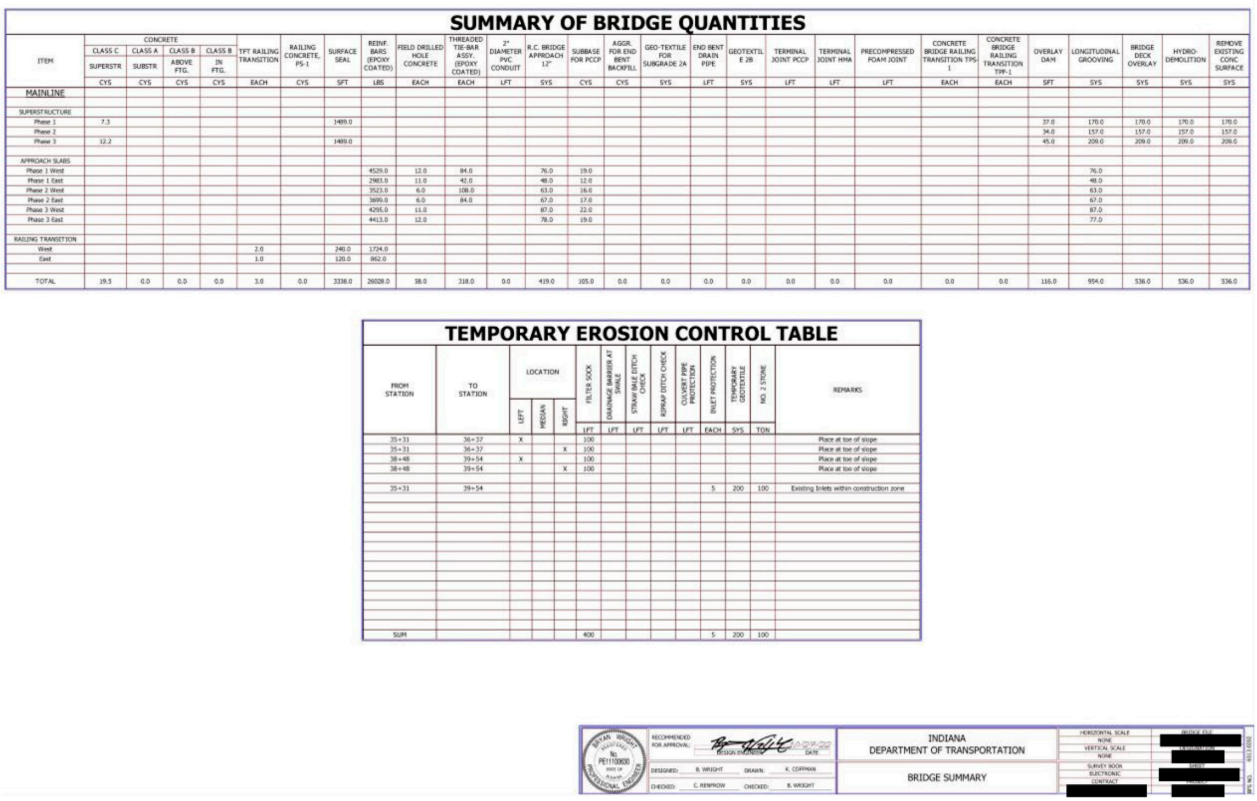


Figure 4.14 Bottom-Region OCR Crop Used to Identify a Target Summary Page and Retain Designation Information.



SUMMARY OF BRIDGE QUANTITIES

ITEM	CONCRETE				TFT RAILING TRANSITION	RAILING CONCRETE, PS-1	SURFACE SEAL	REINF. BARS (EPOXY COATED)	FIELD DRILLED HOLE CONCRETE	THREADED TIE-BAR ASSY. (EPOXY COATED)	2" DIAMETER PVC CONDUIT	R.C. BRIDGE APPROACH 12"	SUBBASE FOR PCCP	AGGR. FOR END BENT BACKFILL	GEO-TEXTILE FOR SUBGRADE 2A	END BENT DRAIN PIPE	GEOTEXTILE 2B	TERMINAL JOINT PCCP	TERMINAL JOINT HMA	PRECOMPRESSED FOAM JOINT	CONCRETE BRIDGE RAILING TRANSITION TYPE I	CONCRETE BRIDGE RAILING TRANSITION TYPE-1	OVERLAY DAM	LONGITUDINAL GROOVING	BRIDGE DECK OVERLAY	HYDRO-DEMOLITION	REMOVE EXISTING CONC SURFACE
	SUPERSTR	SUBSTR	CLASS A ABOVE FTG.	CLASS B IN FTG.																								
	CYS	CYS	CYS	CYS	EACH	CYS	SFT	LBS	EACH	EACH	LFT	SYS	CYS	CYS	SYS	LFT	SYS	LFT	LFT	LFT	EACH	EACH	SFT	SYS	SYS	SYS	SYS
MAINLINE																												
SUPERSTRUCTURE																												
Phase 1	7.3						1489.0																37.0	170.0	170.0	170.0	170.0
Phase 2																							34.0	157.0	157.0	157.0	157.0
Phase 3	12.2						1489.0																45.0	209.0	209.0	209.0	209.0
APPROACH SLABS																												
Phase 1 West								4529.0	12.0	84.0		76.0	19.0											76.0			
Phase 1 East								2983.0	11.0	42.0		48.0	12.0											48.0			
Phase 2 West								3523.0	6.0	108.0		63.0	16.0											63.0			
Phase 2 East								3699.0	6.0	84.0		67.0	17.0											67.0			
Phase 3 West								4295.0	11.0			87.0	22.0											87.0			
Phase 3 East								4413.0	12.0			78.0	19.0											77.0			
RAILING TRANSITION																												
West					2.0		240.0	1724.0																				
East					1.0		120.0	862.0																				
TOTAL	19.5	0.0	0.0	0.0	3.0	0.0	3338.0	26028.0	38.0	318.0	0.0	419.0	105.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	116.0	954.0	536.0	536.0	536.0

TEMPORARY EROSION CONTROL TABLE

FROM STATION	TO STATION	LOCATION			FILTER SOCK	DRAINAGE BARRIER AT SWALE	STRAW BALE DITCH CHECK	RIPRAP DITCH CHECK	CULVERT PIPE PROTECTION	INLET PROTECTION	TEMPORARY GEOTEXTILE	NO. 2 STONE	REMARKS
		LEFT	MEDIAN	RIGHT	LFT	LFT	LFT	LFT	LFT	EACH	SYS	TON	
35+31	36+37	X			100								Place at toe of slope
35+31	36+37			X	100								Place at toe of slope
38+48	39+54	X			100								Place at toe of slope
38+48	39+54			X	100								Place at toe of slope
35+31	39+54									5	200	100	Existing Inlets within construction zone
SUM					400					5	200	100	

Figure 4.15 Example Detected Bridge Plan Table Boundaries and Cells.

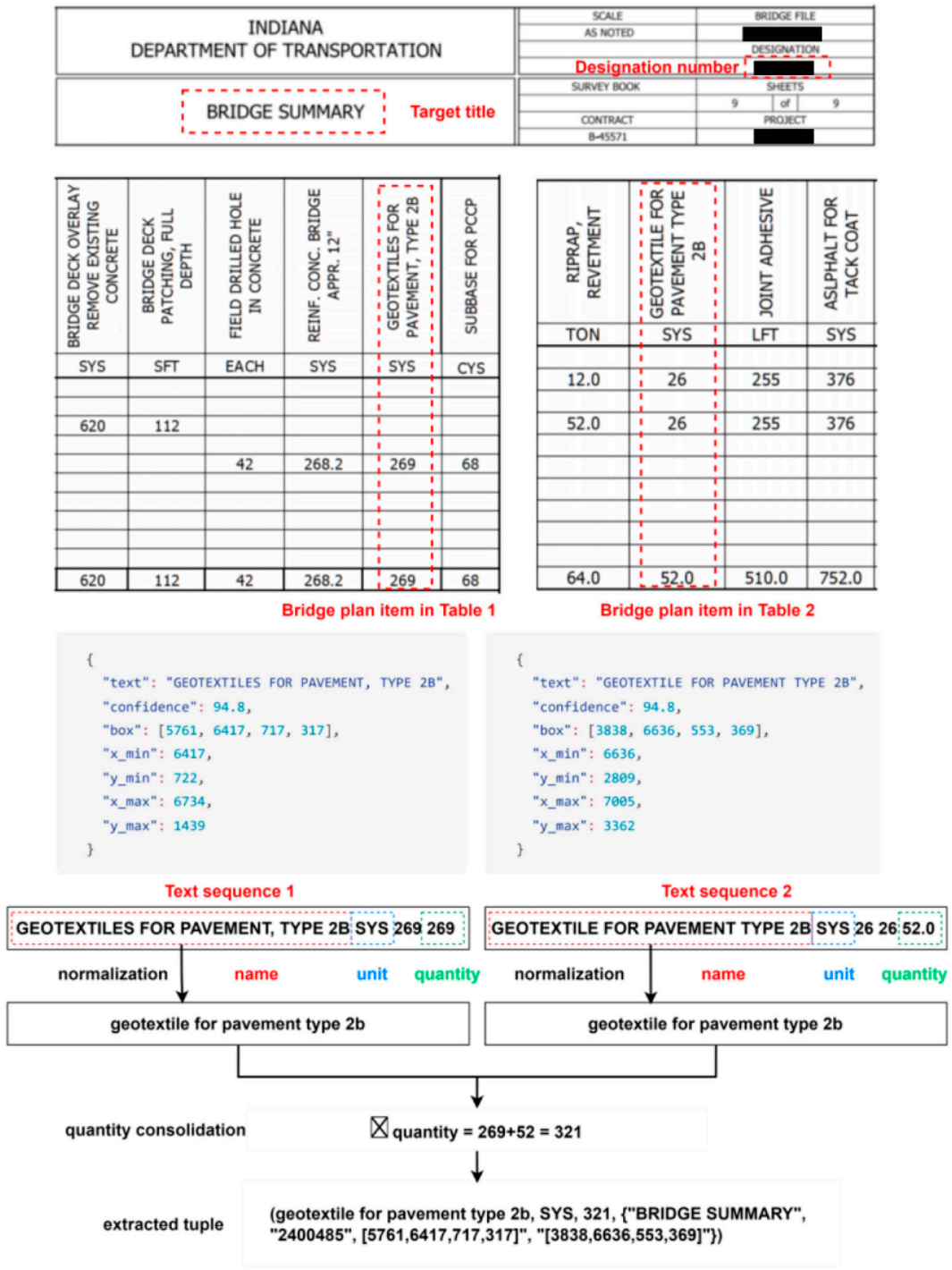


Figure 4.16 Organized OCR Text Sequence and Corresponding Bridge Plan Item Tuple.

4.3.5 Hybrid Matching and Consistency Checking

After extraction, bridge plan items and cost estimate items are grouped by designation number. For each bridge item, the algorithm searches only within the aligned cost estimate group. This prevents items from one project designation from being matched to another designation when a bridge plan or cost estimate package contains multiple project numbers.

The matching algorithm uses a hybrid lexical strategy. Let n_b and n_c denote the normalized bridge-plan and cost-estimate item names. Normalization lowercases the text, removes punctuation, converts hyphens to spaces, standardizes selected OCR variants and item-number expressions, expands recurring abbreviations such as conc, reinf., temp, and br, and removes repeated tokens while preserving their first occurrence. For each bridge plan item, the algorithm searches for the cost estimate item with

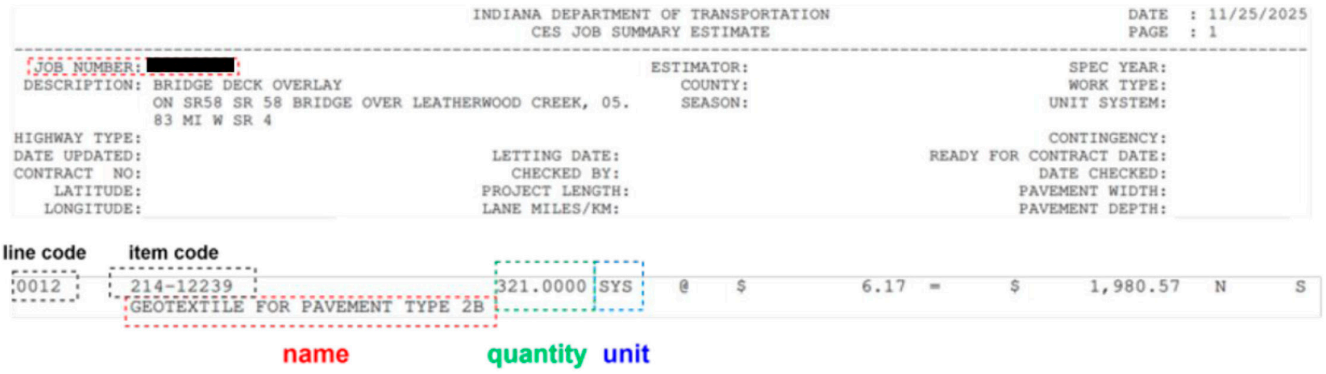


Figure 4.17 Example of Parsing a Cost Estimate Item.

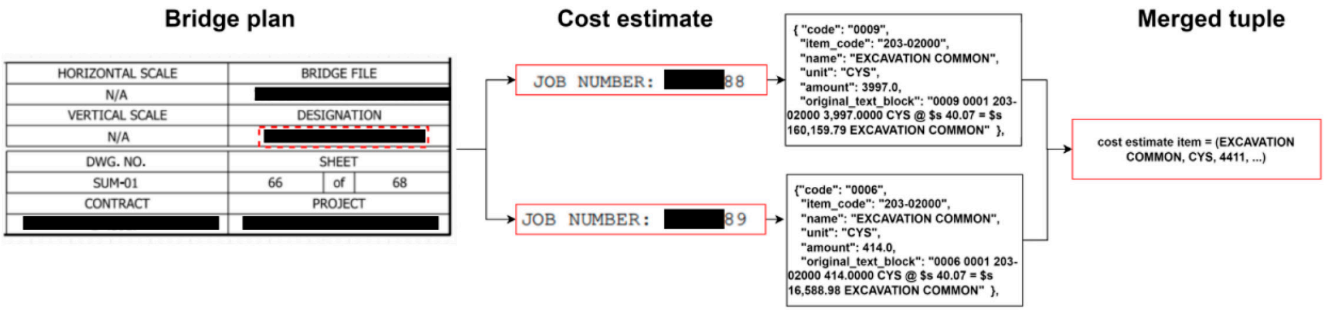


Figure 4.18 Designation-Based Pairing When One Bridge Plan is Associated With Multiple Cost Estimate Files.

TABLE 4.9
Example Part 2 Matching and Checking Records.

Bridge Plan Item	Matched Cost Estimate Item	Checking Result	Note
Geotextile for pavement type 2b, sys, 321	Geotextile for pavement type 2b, sys, 321	Consistent	Exact match
Hydro-demolition; sys; 218	Hydrodemolition; sys; 218	Consistent	Lexical match despite spacing and hyphen
Long. grooving, sys, 218	Longitudinal grooving, sys, 218	Consistent	Match after domain-specific normalization
HMA material for asphalt for tack coat, tons, 605	Asphalt for tack coat, sys, 605	Inconsistent	Lexical match but different units
Surface seal, sft, 656	Surface seal str, ls, 1	Inconsistent	Lexical match but different units and quantities
Length, ft, 58	-	Not found	Design information irrelevant to payment
-	Construction engineering, ls, 1	Additional	Cost estimate item not mapped with any bridge plan item

the highest similarity score. The similarity score is evaluated through a hybrid matching strategy.

The post-processing layer is implemented as an auditable set of rule dictionaries developed from recurring review cases. Rules are applied in a fixed order: records in the forced-not-found dictionary are reported without fuzzy matching; correct-pair rules override lexical scoring by selecting a cost item that contains the mapped target pattern; deny-pair rules remove known incorrect candidates before scoring; candidates containing general-reference terms such as require, see, ref, reference, as per, or note receive a 0.90 score penalty; and duplicate cost candidates with the same normalized name are resolved by preferring the candidate with the same normalized unit. This structure keeps the hybrid matcher data-driven while preserving the reason for each engineering-specific adjustment.

Once the best acceptable cost estimate candidate is selected, consistency checking is deterministic. If normalized units and quantities match, the item is labeled “Consistent.” If the name match is acceptable but either the unit or quantity differs, the item is labeled “Inconsistent.” If no acceptable cost estimate candidate is found, the bridge plan item is labeled “Not Found.” Cost estimate items not mapped to any bridge plan item are recorded as “Additional” cost estimate items for reviewer awareness. Table 4.9 lists some representative consistency checking results.

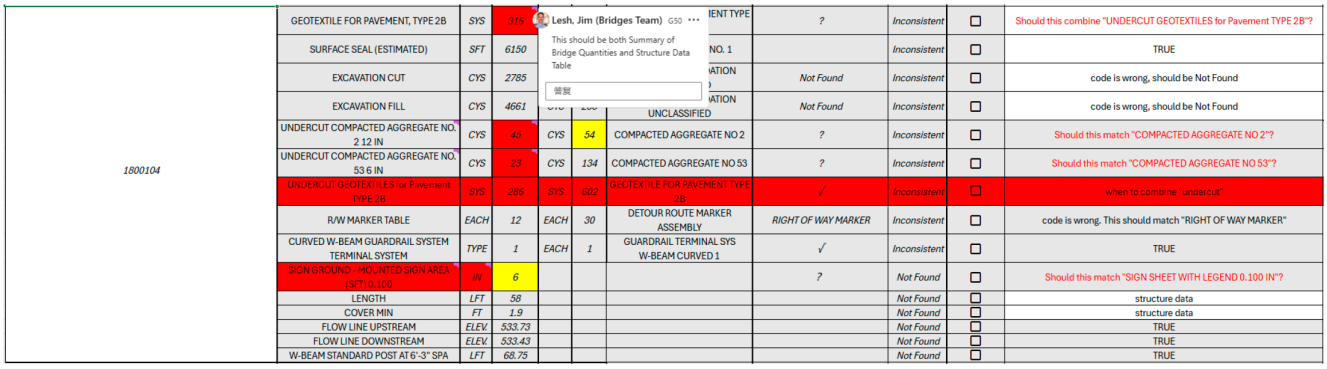


Figure 4.19 Shared Review Spreadsheet Used to Summarize Algorithm Results and Review Comments.

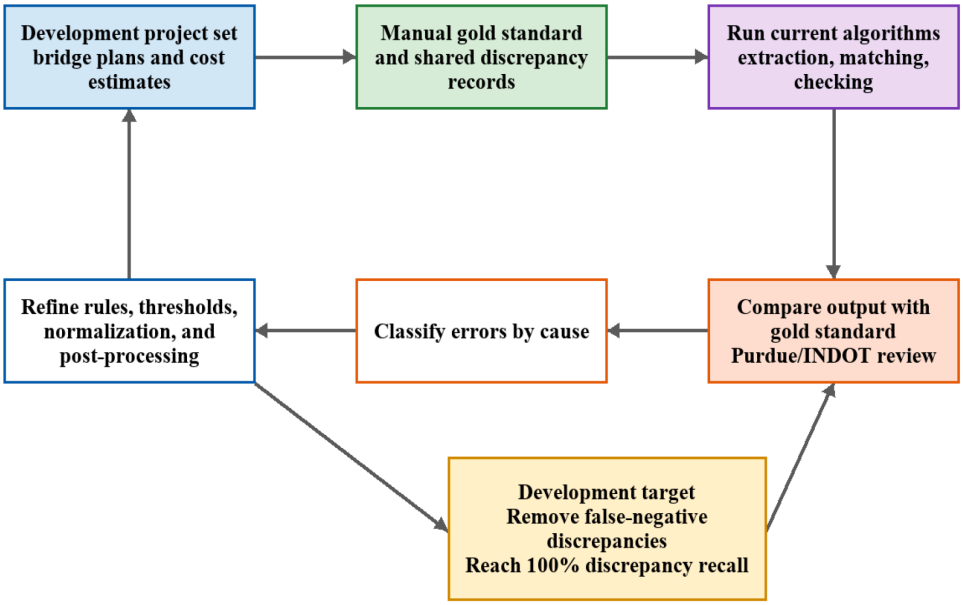


Figure 4.20 Data-Driven Algorithm Development and Purdue/INDOT Review Loop.

4.3.6 Data-Driven Algorithm Refinement

The algorithms supporting comparative checking were developed through an iterative data-driven process. The project team first identified a development set of projects and manually reviewed the bridge plan and cost estimate pay items to produce gold-standard discrepancy records. The results were maintained in a shared spreadsheet so that the algorithm outputs and manual review comments could be compared during recurring technical discussions. Figure 4.19 shows a screenshot of the shared review spreadsheet.

The development goal was to eliminate false negative discrepancies on the development set. In practical terms, the validator was refined until all manually identified inconsistent or not-found bridge plan items were reported.

The development process was organized around three refinement tracks. The first track focused on bridge plan item extraction. The second track focused on cost estimate item extraction. The third track focused on hybrid matching and consistency checking. The overall workflow of refinement is shown in Figure 4.20. Table 4.10 summarizes the dataset during refinement.

TABLE 4.10
Development and Test Data Used for Part 2 Methodology Development.

Dataset Role	Project Designations or Files
Development set for comparative checking	16 project designation groups
Development set for bridge plan extraction	23 bridge summary pages from the development designation groups
Development set for cost estimate extraction	18 cost estimate files associated with the development designation groups
Test set for Part 2 checking	16 project designation groups

Metrics were applied to evaluate the algorithm performance on the development set during refinement. For item extraction, precision measures the proportion of extracted item tuples that exactly matched the gold standard, while recall measures whether gold-standard items were found, including records with field-level mismatches that could still be traced for review. For hybrid matching, strict metrics evaluate complete bridge-cost alignment and checking labels, while discrepancy metrics

TABLE 4.11
Development-Set Performance Changes for Bridge Plan Item Extraction.

Round	Precision	Recall	F1
1	74.29%	75.22%	74.75%
2	80.60%	84.23%	82.37%
3	78.50%	88.82%	83.34%
4	79.12%	89.70%	84.08%
5	87.59%	95.10%	91.19%
6	92.24%	100.00%	95.96%

TABLE 4.12
Development-Set Performance of Cost Estimate Item Extraction.

Development Checkpoint	Precision	Recall	F1	Evaluated Items
Final refined parser	100.00%	100.00%	100.00%	1,160

TABLE 4.13
Development-Set Performance Changes for Hybrid Matching and Consistency Checking.

Round	Strict P/R/F1	Discrepancy P/R/F1
1	82.60% / 87.60% / 85.00%	69.84% / 93.62% / 80.00%
2	83.40% / 89.50% / 86.30%	70.70% / 94.47% / 80.87%
3	84.90% / 91.10% / 87.90%	73.25% / 100.00% / 84.56%
4	86.10% / 92.30% / 89.10%	72.93% / 99.57% / 84.19%
5	91.90% / 98.50% / 95.10%	79.04% / 100.00% / 88.29%

evaluate the engineering objective of identifying potential discrepancies that require manual review.

As shown in Table 4.11, for bridge plan item extraction, recall improved from 75.22% in the first development round to 100.00% in the final round. This result is important because missed bridge-plan items would directly limit the validator's ability to find downstream discrepancies. The final precision remained below 100% because some extra records and field-level mismatches were still generated from OCR and table-structure limitations.

For cost estimate item extraction, the final development checkpoint reached 100% precision and recall, as shown in Table 4.12. Unlike bridge plan tables, the cost estimate files followed a more standardized text structure; after OCR cleanup, item-block segmentation, and regular-expression parsing were refined, all 1,160 development items were recovered correctly.

Table 4.13 summarizes the performance of hybrid matching across development rounds. For hybrid matching and consistency checking, strict F1 increased from 85.00% to 95.10%, indicating improved item alignment and label assignment. More importantly for engineering QA, discrepancy recall reached 100.00% in the final development round, meaning that no manually confirmed discrepancy in the development set was missed. The remaining discrepancy precision errors were false alarms carried forward for reviewer dismissal rather than unreported inconsistencies. Table 4.14 summarizes the observed error patterns and corresponding refinement during development.

4.4 Web Application Integration and Deployment

The final project workflow was integrated into a web application so that the algorithms could be executed as a review-assistance tool. Figure 4.21 illustrates the main web interface, which allows users to select an application based on their required task. After selecting the Drawing Compliance Checker application, the user is greeted with the interface shown in Figure 4.22. On the left grey sidebar is the menu bar, which allows the user to return to the previous main interface (Figure 4.21), view previous task reports, and so on. To start the compliance checker, the user begins by clicking the red "Start New Check" button. The user can then upload one or multiple drawings by browsing or dragging files into the upload area (Figure 4.23). Next, the user clicks on "Validate PDF Size" to check whether the uploaded drawings meet the standard INDOT drawing size requirements (Figure 4.24). Finally, the user clicks on "Start Compliance Check" to proceed with the compliance checking process (Figure 4.25). Figure 4.26 shows the compliance checking report generated once the process is complete. To redo the process, the user needs to scroll back to the top and restart by clicking on "Start New Check."

TABLE 4.14
Examples of Data-Driven Refinements by Algorithm Component.

Component	Observed Error Pattern	Refinement Used in the Final Workflow
Bridge plan extraction	Summary pages missed, rotated text merged incorrectly, unit tokens removed by confidence filtering, and non-payable text extracted as items	Expanded target-title set; added dual-orientation OCR and coordinate transformation; preserved critical unit tokens; refined column grouping and title-block filters.
Cost estimate extraction	OCR lines included headers/footers or failed to separate item blocks when line codes were misread	Added content-start markers, footer stop keywords, tolerant line-code regular expressions, and multi-page line stitching.
Hybrid matching and checking	Different terminology, abbreviation variation, competing cost estimate candidates, and high-overlap false matches	Expanded normalization dictionaries; selected unit-matching candidates; added correct-pair, deny-pair, and forced-not-found rules.
Report generation	Reviewers needed to distinguish confirmed discrepancies from algorithm false alarms	Preserved bridge evidence, cost estimate evidence, matched fields, issue notes, and downloadable CSV/JSON/text outputs.

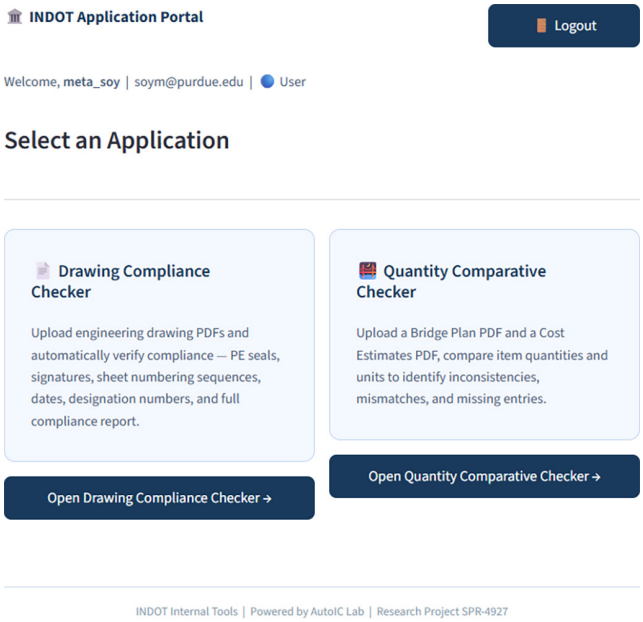


Figure 4.21 Application Interface.

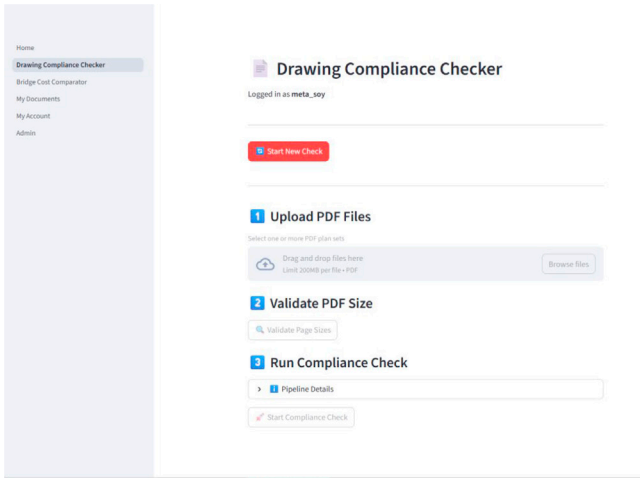


Figure 4.22 Bridge Plan Formatting Compliance Checker Interface.

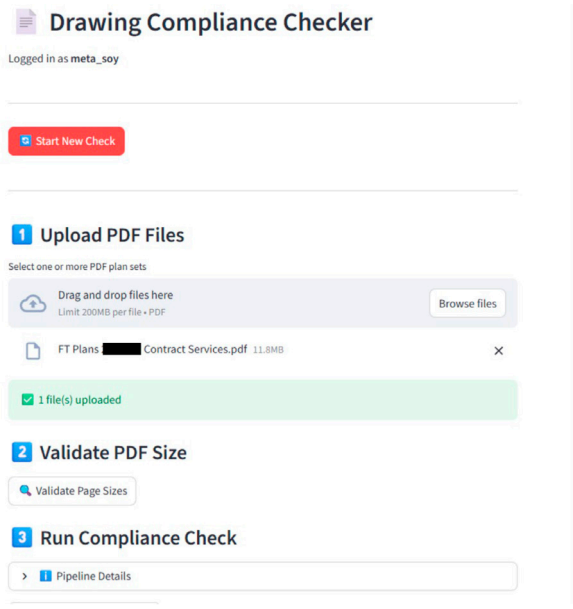


Figure 4.23 PDF Upload Step for Compliance Checking.

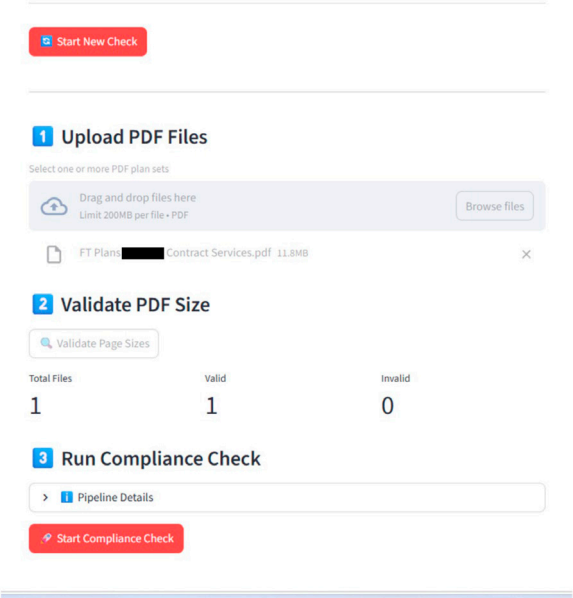


Figure 4.24 PDF Size Validation Before Compliance Checking.

For comparative checking, the application accepts bridge plan and cost estimate PDFs, extracts designation numbers from filenames, pairs bridge plan files with the expected cost estimate files and organizes outputs under project-specific job folders. When a bridge plan contains more than one designation number, the workflow combines the corresponding cost estimate item sets before comparison. Figure 4.27 shows the interface of uploading deliverables.

As shown in Figure 4.28, the web implementation preserves intermediate outputs, including target-page scan crops, page-level bridge OCR files, organized bridge item files, structured cost estimate files, comparison JSON files, CSV reports, and text summaries. Figure 4.29 shows the difference table, and the full outputs can be downloaded through the interface shown in Figure 4.30. This output organization is important for implementation because it allows an engineer or developer to trace a flagged issue back to the page crop, OCR text, structured item tuple, and matching decision that produced it. Figure 4.31 summarizes the web application workflow. Table 4.15 summarizes the output artifact list.

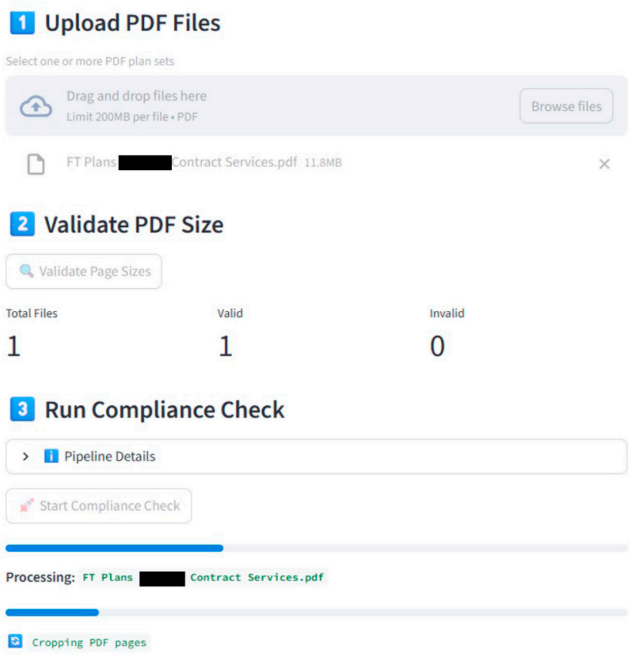


Figure 4.25 Running the Compliance Checking Process.

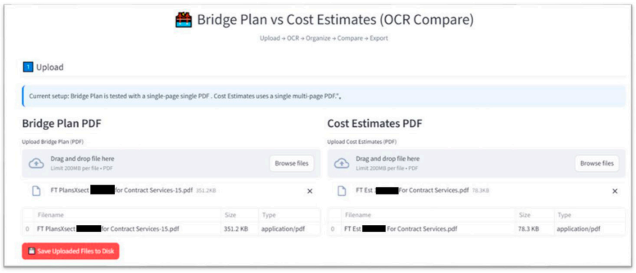


Figure 4.27 Comparative Checking Upload Interface.

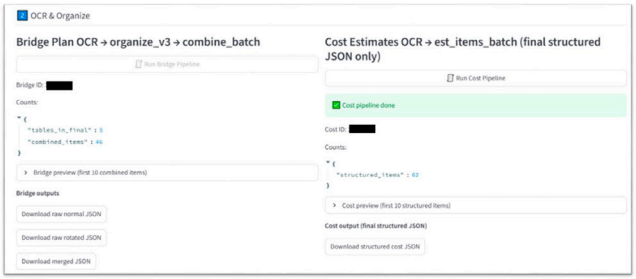


Figure 4.28 Interface of Comparative Checking Intermediate Outputs.

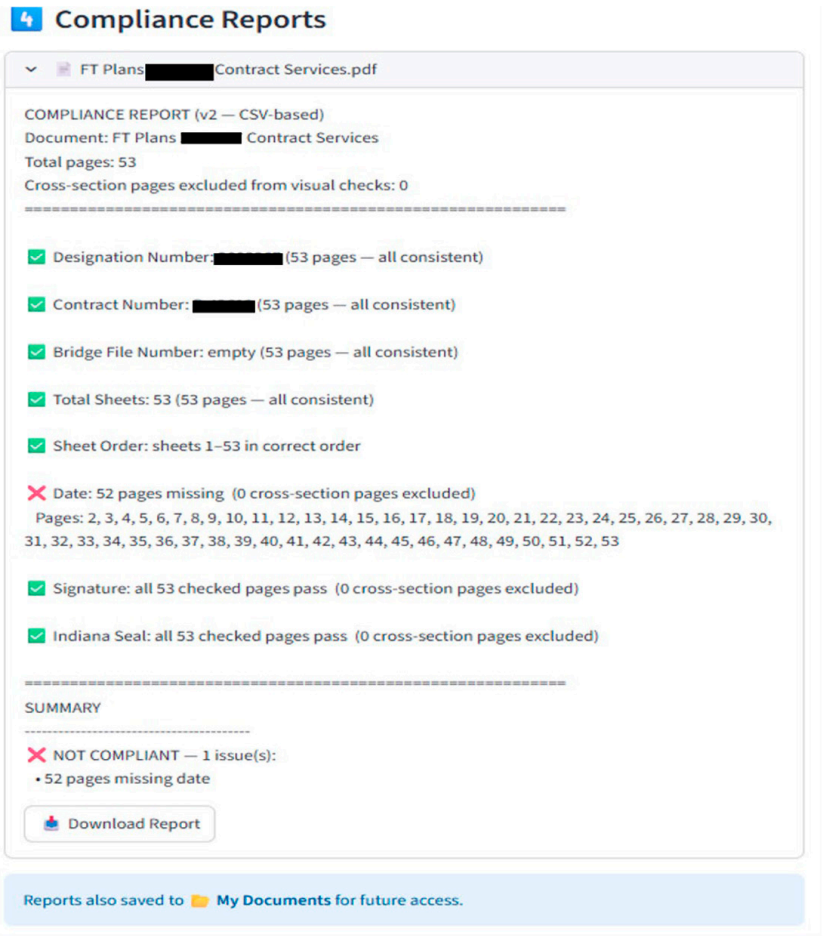
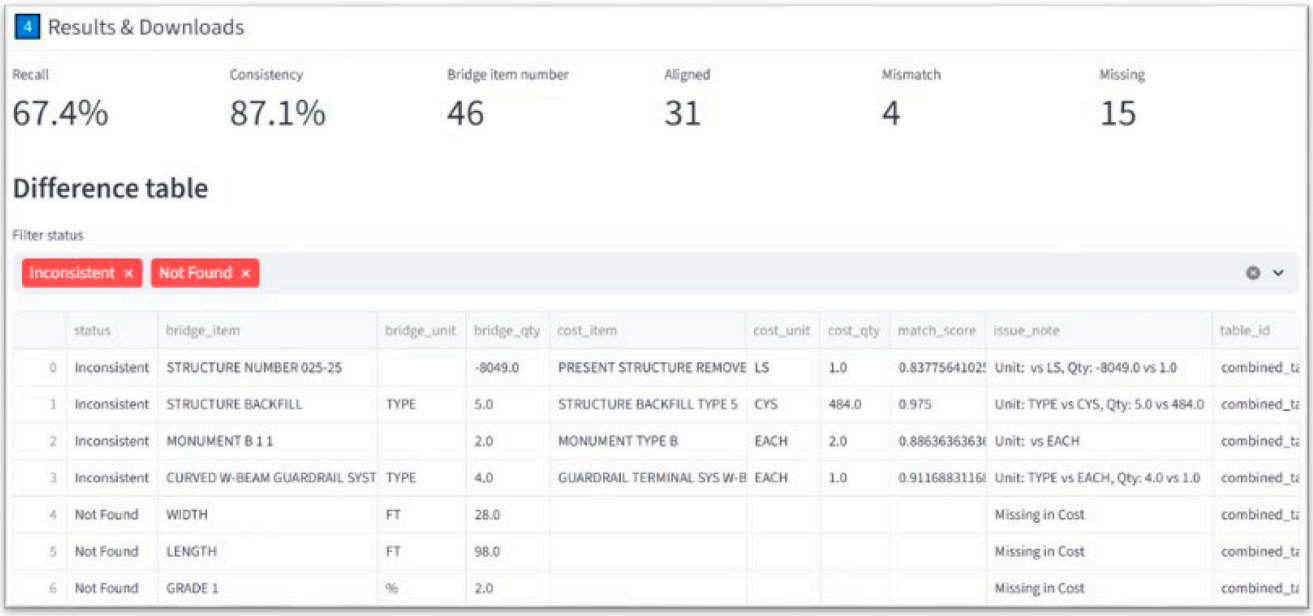


Figure 4.26 Compliance Checking Report Output.



	status	bridge_item	bridge_unit	bridge_qty	cost_item	cost_unit	cost_qty	match_score	issue_note	table_id
0	Inconsistent	STRUCTURE NUMBER 025-25		-8049.0	PRESENT STRUCTURE REMOVE	LS	1.0	0.83775641025	Unit: vs LS, Qty: -8049.0 vs 1.0	combined_ta
1	Inconsistent	STRUCTURE BACKFILL	TYPE	5.0	STRUCTURE BACKFILL TYPE 5	CYS	484.0	0.975	Unit: TYPE vs CYS, Qty: 5.0 vs 484.0	combined_ta
2	Inconsistent	MONUMENT B 1 1		2.0	MONUMENT TYPE B	EACH	2.0	0.886363636363	Unit: vs EACH	combined_ta
3	Inconsistent	CURVED W-BEAM GUARDRAIL SYST	TYPE	4.0	GUARDRAIL TERMINAL SYS W-B	EACH	1.0	0.911688311688	Unit: TYPE vs EACH, Qty: 4.0 vs 1.0	combined_ta
4	Not Found	WIDTH	FT	28.0					Missing in Cost	combined_ta
5	Not Found	LENGTH	FT	98.0					Missing in Cost	combined_ta
6	Not Found	GRADE 1	%	2.0					Missing in Cost	combined_ta

Figure 4.29 Example of Difference Table.

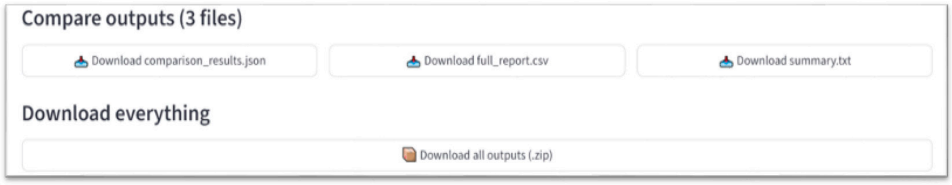


Figure 4.30 Interface of Downloading Comparative Checking Output Files.

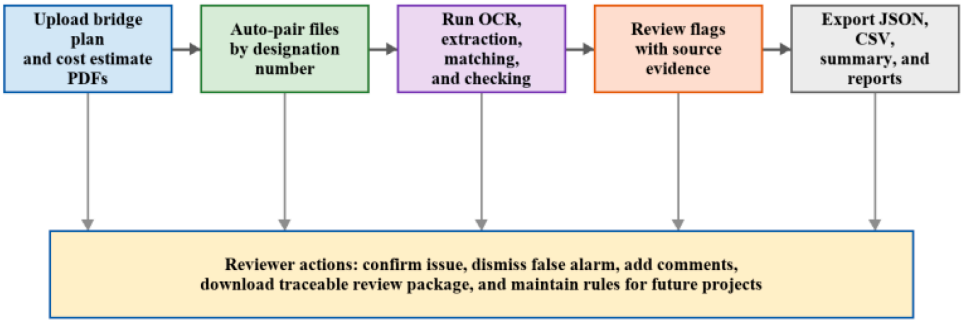


Figure 4.31 Web Application Workflow for Project-Level Validation and Reviewer Interaction.

TABLE 4.15
Main Output Artifacts Preserved by the Workflow.

Output Artifact	Example Content	Review or Maintenance Use
Target-page scan output	Matched page list and bottom-right crop OCR text	Verify whether the correct summary pages were selected.
Bridge organized/final items	Item name, unit, quantity, table ID, and source page metadata	Check bridge extraction and within-plan combining decisions.
Structured cost estimate items	Line code, item code, item name, unit, amount, and original OCR text block	Check cost estimate parsing and source evidence.
Comparison report	Status, bridge item, matched cost item, unit/quantity fields, match score, and issue note	Support engineering review of discrepancies and false alarms.

5. IMPLEMENTATION CONSIDERATIONS

5.1 Review-Assistance Use

The Bridge Data Virtual Validator should be implemented as a review-assistance system rather than an automatic approval system. The intended workflow is to prescreen Final Tracings deliverables; generate traceable flags; and allow INDOT reviewers to confirm, dismiss, or comment on each issue. This use case is important because the current algorithms are designed to preserve high recall for potential discrepancies while relying on reviewer judgment to resolve false alarms, scope differences, or project-specific exceptions.

The reviewer-facing output should keep compliance flags separate from pay item discrepancy flags. Compliance checking answers whether required bridge plan letting control information is present and internally consistent across the plan set. Comparative checking answers whether pay item information agrees between bridge plans and cost estimate. Separating the two result types will make the review process easier to assign, document, and improve over time.

5.2 Process and Data Integration

Implementation should begin with a pilot workflow using a defined group of Final Tracings deliverables. For each project, the validator can be used to upload bridge plan and cost estimate files, run the compliance and comparative checking modules, download the review package, and record reviewer actions. Confirmed issues can then be routed back to the appropriate project QA process before letting or before downstream data entry.

The project identified a longer-term goal of using validated Final Tracings information to support iTAMS data entry and the creation of a bridge birth certificate. The current prototype supports this direction by converting document evidence into structured and traceable records. Additional implementation work will be needed to define which validated fields should be transferred, how reviewer approval should be recorded, and how the resulting data should be synchronized with INDOT business systems.

5.3 Algorithm Maintenance

The validator should be continuously maintained. OCR behavior, bridge plan formats, cost estimate wording, and agency terminology may change across projects and over time. Therefore, compliance definitions, target titles, and post-processing rules should be versioned and managed as editable configuration resources where possible. Reviewer feedback should be stored in a structured format so that confirmed false alarms and missed issues can be used to refine the underlying algorithms.

6. SUMMARY AND RECOMMENDATIONS

6.1 Summary and Findings

This project developed the foundation for Bridge Data Virtual Validator, a tool intended to support automated QA of INDOT bridge project deliverables. The project responded to the practical need to review Final Tracings information that is distributed across bridge plans, cost estimates, and other submitted documents. Rather than replacing current deliverable formats, the project demonstrated how selected information can be extracted from existing PDFs, structured into validation records, and checked with traceable evidence.

Two complementary validation functions were implemented within the Bridge Data Virtual Validator. The bridge plan format compliance checking module checks page-bottom title-block information, including file number, signature, date, stamp, and page number, and evaluates whether required information is present and consistent across a plan set. The cross-document comparative checking module evaluates whether pay items listed in bridge plan summary tables are consistent with itemized cost estimate records for the same project designation.

For bridge plan format compliance checking, the project developed an OCR-, computer-vision-, and rule-based workflow that extracts title-block evidence, detects visual elements such as seals and signatures, evaluates page and file-number consistency, and generates traceable noncompliance reports. For pay-item comparative checking, the project developed an end-to-end workflow that screens bridge summary pages, extracts pay items from plan tables and cost estimates, aligns designation numbers, applies normalization and hybrid lexical/rule-based matching, and reports item-level consistency, inconsistency, and not-found conditions.

A web application prototype was also developed to package the workflow for practical use. The application supports file upload, designation-based pairing, intermediate output storage, comparison report generation, and downloadable review artifacts. This prototype shows how the validator can move from research scripts toward a review-assistance tool for agency use.

The project demonstrates the feasibility of converting existing PDF-based Final Tracings deliverables into structured validation records without requiring changes to current submission formats. Across both validation functions, the emphasis is on source evidence, explainable checking logic, and human-in-the-loop review. The resulting web application and report outputs provide a practical foundation for pilot deployment, reviewer feedback collection, algorithm maintenance, and future expansion to additional bridge data objects.

6.2 Recommendations and Implementation

The following implementation steps are recommended. First, the bridge plan format compliance checking and cross-document comparative checking functions should be maintained as complementary modules within one project-level review application. Compliance flags and pay-item discrepancy flags should remain separated in the interface and reports so that reviewers can assign, interpret, and resolve each type of issue appropriately. Second, reviewer decision fields should be added to the application so that each reported item can be marked as a confirmed issue, false alarm, resolved issue, or item requiring follow-up. These decisions should be stored with source evidence and reviewer comments. Third,

the Bridge Data Virtual Validator should be piloted on additional Final Tracings projects representing different bridge types, project types, consultants, document formats, and scan qualities. Pilot review should track not only algorithm performance, but also reviewer effort, confirmed issue types, false alarm patterns, and the time required to resolve each report. Fourth, target-page title lists, OCR correction rules, normalization dictionaries, matching rules, and compliance definitions should be maintained as versioned resources so that the tool can be updated as more projects are reviewed and documentation practices evolve.

Future implementation could also expand the validation scope. Quantity calculations and BrR-related information could be evaluated as future data sources once the plan and cost estimate workflow is stable. The longer-term implementation pathway should connect validated Final Tracings records to iTAMS and bridge birth certificate workflows after data ownership, approval status, and transfer requirements are defined.

6.3 Expected Benefits and Future Work

The expected benefits of the Bridge Data Virtual Validator include reduced manual effort in locating potential document issues, more efficient review of Final Tracings deliverables, earlier identification of pay-item discrepancies, improved traceability between review comments and source documents, and a foundational support for downstream digital delivery and data-management workflows. By identifying discrepancies before they propagate into construction or payment processes, the validator has the potential to reduce review risk and support more reliable bridge project records.

Future work should improve OCR robustness for dense tables and low-quality scans, strengthen extraction for unusual plan formats and special symbols, add reviewer decision logging, evaluate additional cost estimate items as a separate review category, and expand the validator to other deliverables and data objects. Continued pilot use and reviewer feedback will be essential for converting the current prototype into a full-blown agency QA tool.

REFERENCES

Akanbi, T., & Zhang, J. (2022). Framework for developing IFC-based 3D documentation from 2D bridge drawings. *Journal of Computing in Civil Engineering*, *36*(1). https://doi.org/10.1061/(ASCE)CP.1943-5487.0000986

Du, Y., Li, C., Guo, R., Yin, X., Liu, W., Zhou, J., Bai, Y., Yu, Z., Yang, Y., Dang, Q., & Wang, H. (2020). *PP-OCR: A practical ultra lightweight OCR system*. arXiv. https://doi.org/10.48550/arXiv.2009.09941

Federal Highway Administration. (2025). *SNBI data validation logic: Version 1.1*. https://www.fhwa.dot.gov/bridge/snbi/data_validation_logic.cfm

Indiana Department of Transportation. (2024). *Final TRACINGS submittal documents and project data for SPR-4927: Bridge data virtual validator* [Project documentation]. Indiana Department of Transportation.

Li, H., & Zhang, J. (2025). Semiautomatic IFC-based BIM reconstruction for as-designed bridges from 2D Plans leveraging semantic segmentation and enrichment. *Journal of Computing in Civil Engineering*, *39*(6). https://doi.org/10.1061/JCCEE5.CPENG-5931

New York State Department of Transportation. (2021). Bridge manual. https://www.dot.ny.gov/divisions/engineering/structures/repository/manuals/brman-usc/NYSDOT_Bridge_Manual_2021.pdf

Shafei, B. (2024). *Automated data feature extraction from bridge plans* (NCHRP IDEA 20-30/IDEA 230). Transportation Research Board. https://onlinepubs.trb.org/onlinepubs/IDEA/FinalReports/Highway/NCHRP230.pdf

Soy, M., & Zhang, J. (2025). GPT4-vision multimodal model-powered query-answering chatbot for bridge PDF drawings. In B. Akinci, M. Bergés, F. Jazizadeh, C. C. Menassa, & J. Yeoh (Eds.), *Computing in civil engineering 2024: Artificial intelligence, automation and robotics, and human-centered innovations* (pp. 505–513). American Society of Civil Engineers. https://doi.org/10.1061/9780784486115.053

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2022). *YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors*. arXiv. https://doi.org/10.48550/arXiv.2207.02696

Zhang, J., Li, H., Akanbi, T., Hegazy, H., Yang, F., Soy, M., Chen, Y., & Dib, H. N. (2026). *Integration of BIM for bridge modeling* (Joint Transportation Research Program Publication No. FHWA/IN/JTRP-2026/08). Purdue University. https://doi.org/10.5703/1288284318616

About the Joint Transportation Research Program (JTRP)

On March 11, 1937, the Indiana Legislature passed an act which authorized the Indiana State Highway Commission to cooperate with and assist Purdue University in developing the best methods of improving and maintaining the highways of the state and the respective counties thereof. That collaborative effort was called the Joint Highway Research Project (JHRP). In 1997 the collaborative venture was renamed as the Joint Transportation Research Program (JTRP) to reflect the state and national efforts to integrate the management and operation of various transportation modes.

The first studies of JHRP were concerned with Test Road No. 1—evaluation of the weathering characteristics of stabilized materials. After World War II, the JHRP program grew substantially and was regularly producing technical reports. Over 1,600 technical reports are now available, published as part of the JHRP and subsequently JTRP collaborative venture between Purdue University and what is now the Indiana Department of Transportation.

Free online access to all reports is provided through a unique collaboration between JTRP and Purdue Libraries. These are available at docs.lib.purdue.edu/jtrp/.

Further information about JTRP and its current research program is available at engineering.purdue.edu/JTRP.

About This Report

An open access version of this publication is available online. See the URL in the citation below.

Zhang, J., Pang, B., Soy, M., Nartey, B., & Chen, Y. (2026). *Bridge data virtual validator* (Joint Transportation Research Program Publication No. FHWA/IN/JTRP-2026/23). West Lafayette, IN: Purdue University. https://doi.org/10.5703/1288284318701