AI-Powered PDF to Accessible HTML Conversion System

University of Illinois Chicago - Technical Implementation Proposal

Executive Summary

The University of Illinois Chicago has the opportunity to become the first major research institution to deploy artificial intelligence as an "epistemological translator"—transforming how academic knowledge reaches all learners through accessible digital formats. This proposal outlines a comprehensive technical system that converts PDF documents into fully accessible HTML using transparent AI processing, creating permanent institutional assets while addressing critical accessibility compliance requirements.

Core Innovation: Unlike commercial black-box solutions, this system provides transparent AI reasoning that educators can review and correct, ensuring accurate knowledge transfer across diverse learning modalities. The approach bridges the gap between creator intent and inclusive user experiences through multi-agent AI processing with human oversight capabilities.

Technical Foundation: Apple's Math Notes demonstrates large-scale visual-to-semantic extraction, while ArXiv's 2M+ paper corpus shows successful academic LaTeX-to-HTML conversion with community acceptance of iterative improvement approaches.

Strategic Value: UIC gains open-source accessibility infrastructure suitable for academic publication, grant positioning, and community contribution, while creating permanent accessible assets with versioned URLs perfect for academic citations and long-term preservation.

Technical Architecture

Multi-Agent Processing Pipeline

The system employs a multi-agent architecture using PydanticAI for coordination and model management. Each agent specializes in different document components, with automatic confidence scoring that adjusts the system confidence when agents detect particular indicators for a given document:

PDF Document → LangChain Processing → Multi-Agent Analysis → Semantic Caches → MDX Generation → Astro Static Site → S3 Hosting

Agent Specialization & Model Selection

Structure Agent: Fast document hierarchy extraction accomplished via low-end language models

- Heading detection and semantic nesting
- Reading order analysis and landmark identification
- Document outline generation with navigation structure
- Table of contents reconstruction with cross-references

Mathematics Agent: Looks out for LaTeX and mathematical representations using specialized mathematical models

- Inline and display equation processing
- Formula structure analysis and accessibility tagging
- Mathematical notation preservation with screen reader compatibility
- Alternative text generation for complex equations

Tables Agent: Precise tabular data relationships and header associations using multi-modal models.

- Cell structure analysis with proper scope identification
- Header detection and association mapping
- Data export capabilities (CSV, JSON) for alternative access via tool use
- Summary statistics generation for complex datasets

Figures Agent: Context-aware visual analysis and description generation using multi-modal models

- Image classification and content analysis
- Caption extraction and contextual enhancement
- Long description generation based on surrounding content
- Chart data extraction for accessible alternatives

Citations Agent: Reference integrity and cross-link preservation using low-end reasoning models to trace multiple threads.

- Citation format detection and structured extraction
- DOI and URL validation with broken link detection
- Cross-reference mapping
- Academic integrity preservation across format conversion

Typography Agent: Semantic meaning extraction from textual formatting and semantics.

- Emphasis detection and semantic tagging
- List structure identification and proper markup
- Font-based meaning interpretation (italic → emphasis, bold → strong)
- Visual hierarchy translation to semantic structure
- Typography usage which changes textual context

Semantic Caching System

Each agent produces human-readable insights stored as semantic caches—observations like "italicized text positioned near photograph, likely caption relationship" or "bold text following numbered sequence, probable heading structure." These cached insights enable:

- **Transparent Reasoning**: Educators can review specific AI observations and reasoning chains
- **Human Validation**: Simple interface for invalidating incorrect assumptions, triggering recalculation
- **Research Corpus**: Semantic insights and corrections form valuable dataset for ML research and model training
- **Iterative Improvement**: Human-validated reasoning improves processing quality across the institution

Technology Stack Rationale

Technology	Purpose	Technical Justification
PydanticAI	Agent coordination and model management	Enables multi-model orchestration with type-safe agent communication, allowing optimal model selection for specialized tasks
LangChain	PDF processing and document loading	Mature PDF parsing with multiple extraction and annotation strategies, handles diverse document formats and quality levels
MDX	Intermediate content format	Plain text portability with component integration capabilities, preserves semantic structure while enabling interactive enhancements
Astro	Static site generation	Optimized performance for document-heavy content, native MDX support, React component integration without client-side JavaScript overhead
Shadcn/Radix	UI component primitives	Accessible-by-default and extendable components built on Radix UI primitives, ensures WCAG 2.1 AA compliance out of the box while allowing for complete code ownership.
Tailwind CSS	Styling and theming system	Extensive customization capabilities enabling user preference adaptation, high-contrast themes, and responsive design

Infrastructure Specifications

AWS Architecture & Deployment

Container Orchestration: AWS ECS (Elastic Container Service) with Fargate for serverless container management

- Minimal resource allocation: 0.25-1 vCPU, 1-4GB RAM for typical document processing
- Docker containerization for consistent deployment across environments
- Pay-per-use pricing model with no idle costs

Queue Management System: Redis-based job processing with priority scheduling

- Concurrent processing: 4-6 documents simultaneously based on complexity
- Priority queues: Academic deadlines, faculty requests, bulk processing
- Dead letter queue handling for failed conversions with retry logic
- Processing status tracking with real-time updates

Storage Infrastructure:

- **AWS S3**: Static site hosting with CloudFront CDN for global accessibility
- **Versioned storage**: Multiple document versions with permanent URLs
- **Backup systems**: Cross-region replication for institutional content preservation
- **Access controls**: Fine-grained permissions for UIC domain management

Simplified Processing Approach:

Document Size	Processing Time	Resource Allocation	Use Cases
1-25 pages, <20MB	2-8 minutes	0.5 vCPU, 2GB RAM	Pilot scope - covers ~70% of UIC corpus
25+ pages (Future)	Variable	Scaled resources	Deferred to extensions

URL Versioning & Academic Citation System

The system implements a comprehensive versioning approach following academic repository patterns:

URL Structure Patterns (example):

Primary: `https://docs.uic.edu/{document-slug}/{version}/`
Latest: `https://docs.uic.edu/{document-slug}/latest/`

Course Context: `https://docs.uic.edu/{course}/{document-slug}/{version}/`
DOI Integration: `https://doi.org/10.25418/uic.{identifier}`

Versioning Features:

- **Permanent identifiers**: Each version receives immutable URL for academic citations
- **Semantic linking**: Metadata connections between versions (not just URL suffixes)
- **DOI minting**: Automatic DOI assignment following Zenodo/Figshare patterns
- **Rollback capability**: Previous versions remain accessible indefinitely
- **UIC branding**: Institution-owned permanent URLs independent of external services

Metadata Preservation:

- Version history with changelog tracking
- Processing confidence scores and human review status
- Original PDF source linking
- Accessibility metrics and compliance reporting

User Experience Design

Faculty/Staff Interface

Document Submission Workflow:

1. **Upload Interface**: Drag-and-drop PDF submission with batch processing capability
2. **Processing Dashboard**: Real-time status updates with confidence scoring visualization
3. **Review Interface**: AI decision transparency with correction capabilities
4. **Publishing Controls**: Version management and URL generation tools

Review & Correction Interface:

AI Decision Transparency: The system presents its structural determinations as conversational explanations rather than technical outputs:

- **Heading Analysis**: "I identified this as a Level 2 heading because of its bold formatting and position after the introduction paragraph"
- **Alt Text Reasoning**: "This image appears to be a flowchart showing three connected processes, so I described the workflow and data relationships"
- **Table Structure**: "I detected this table has merged header cells spanning two columns, creating scope relationships for accessibility"
- **Citation Context**: "This reference links to Figure 3 on page 15, maintaining cross-document navigation"

Natural Language Correction Workflow:

- Faculty can describe adjustments in plain language: "This heading should be Level 3, not Level 2" or "The alt text is missing the key data trend shown in the graph"
- AI processes feedback and attempts automatic corrections with transparent reasoning
- Iterative conversation allows refinement: "Better, but also mention the x-axis represents time periods"
- Faculty sees AI update MDX which is then rendered to see if it matches expectation
- Confidence scoring updates based on human validation and correction frequency

Fallback Manual Controls:

- **Raw MDX Editor**: Direct markdown editing for complex structural changes
- **Semantic Cache Editor**: Advanced users can modify AI reasoning chains directly

API-First Design for System Integration

Core API Endpoints

Document Submission

POST /api/documents/submit

- Accepts PDF files with metadata
- Returns processing job ID and estimated completion time
- Supports batch submission for multiple documents

Processing Status

GET /api/documents/{id}/status

- Real-time processing status with confidence metrics
- Error reporting with specific failure categorization
- Progress tracking with stage-by-stage updates

Document Retrieval

GET /api/documents/{id}/result

- Retrieval of converted HTML and associated assets
- Accessibility compliance reports with WCAG 2.1 AA validation
- Semantic cache access for transparency and corrections

Feedback & Corrections

POST /api/documents/{id}/feedback

- Correction submission for AI decision improvements
- User annotation and override capabilities

Version Management

GET /api/documents/{id}/versions

- Version history and comparison tools
- Rollback capabilities with change tracking
- Citation URL generation for academic reference

Content Detection & User Guidance

Unsupported Content Detection & Processing:

Since faculty can upload any PDF content via drag-and-drop interface, the system processes all documents while providing transparent quality expectations:

PDF Analysis → Content Classification → Confidence Scoring → Processing + Banner Notification

├── High Confidence (>85%) → Standard processing with minimal warnings

├── Medium Confidence (60-85%) → Process with "Review Recommended" banner

└── Low Confidence (<60%) → Process with "Check Output Carefully" banner

Content Detection & User Guidance:

- **Automatic Flagging**: LaTeX equations, complex charts, scanned/OCR-only text detected and flagged
- **Banner System**: Clear visual indicators on converted documents indicating potential issues
- **Processing Philosophy**: Attempt conversion of all content while transparently communicating limitations
- **Faculty Review**: All flagged content includes specific guidance for manual verification

Canvas Integration & LMS Compatibility

Primary Integration Approach

External URL Module Items: UIC-hosted permanent URLs for academic citations

- Institutional content ownership with versioned, citable URLs
- Content optimized for accessibility compliance and screen reader navigation
- Responsive design with mobile and assistive technology prioritization
- Interactive enhancements for improved accessibility (collapsible sections, focus management)

Compatibility Options

Custom Rendering: different rendering requirements supported

- Accessed via query parameter flag (?static=true)
- Print-optimized CSS for physically printing documents.

Equalify Platform Integration

Workflow Integration Points:

1. **Automatic Detection**: Equalify scans identify PDFs with accessibility violations
2. **Processing Trigger**: API webhook initiates conversion workflow automatically
3. **Quality Validation**: Re-scanning of converted HTML provides before/after metrics
4. **Dashboard Integration**: Accessibility improvement tracking within existing Equalify interface
5. **Reporting Enhancement**: Institutional accessibility metrics include conversion success rates

Technical Compatibility:

- Mirror Equalify's PHP/MySQL queue architecture for institutional familiarity
- Integrate with existing `property_id` and `page_url` database structure
- Maintain violation categorization system compatibility
- Minimal disruption to established UIC IT workflows

UIC Corpus Analysis & Pilot Strategy

Data-Driven Document Selection

Based on analysis of UIC's comprehensive PDF audit (3,522 documents), the pilot strategy prioritizes maximum accessibility impact:

Document Distribution Analysis:

- **61.8% are 1-5 pages** (immediate broad impact opportunity)
- **Median length: 3 pages** (validates "simple first" strategy)
- **90th percentile: ~28 pages** (most documents manageable in early phases)
- **Document types**: Course materials (38%), research papers (22%), administrative documents (18%), theses/dissertations (12%), policy documents (10%)

Intelligent Pilot Selection Criteria:

- **Phase 1 Focus**: Text-dominant documents ≤25 pages covering ~70% of corpus
- **Department Diversity**: Representative sampling across STEM, humanities, professional schools
- **Quality Variation**: Mix of tagged/untagged PDFs, scanned/native documents
- **Complexity Progression**: Simple structures → tables → figures → mathematical content

Pilot Document Categories:

1. **Course Materials** (8 documents): Syllabi, assignment sheets, course policies
2. **Research Papers** (7 documents): Journal articles, conference papers, preprints
3. **Administrative Documents** (6 documents): Policy documents, procedure manuals, reports
4. **Technical Content** (5 documents): STEM papers with equations, data tables, figures
5. **Edge Cases** (4 documents): Scanned documents, complex layouts, multi-language content

Success Metrics & Quality Assurance

Automated Validation:

- **WCAG 2.1 AA Compliance**: Equalify and axe-core accessibility scanning
- **Performance Metrics**: Page load times, screen responsiveness and reflow, mobile optimization
- **Cross-Device Testing**: Compatibility across assistive technologies and devices

Manual Quality Review:

- **Screen Reader Testing**: Windows+NVDA+Chrome, Android+TalkBack+Chrome, Mac+Safari+VoiceOver, iPhone+Safari+VoiceOver
- **Keyboard Navigation**: Complete functionality without mouse interaction
- **UIC Accessibility Team Approval**: Institutional standards verification for converted content

Academic Compliance Metrics:

- **Structure Accuracy**: ≥90% proper heading hierarchy preservation
- **Image Accessibility**: 100% images tagged as decorative or provided with meaningful alt text
- **Processing Reliability**: Zero critical failures with clear error handling
- **Review Efficiency**: ≤10 minutes faculty review time for typical 10-page document

Foundation System & Future Extensions

Foundation Infrastructure ($60,000)

Core System Development:

- Multi-agent PydanticAI pipeline with semantic caching
- AWS ECS deployment with Redis queue management
- Basic web interface for document submission and review
- S3 static hosting with versioned URL generation
- API endpoints for external integration

Pilot Validation:

- 30 documents processed across complexity tiers
- Comprehensive accessibility compliance reporting
- Faculty review interface testing and refinement
- Integration documentation for Equalify and Canvas
- Performance benchmarking and optimization

Deliverables:

- Working PDF-to-HTML conversion system
- Transparent AI processing with educator review capabilities
- Professional accessibility compliance reports
- Technical documentation for system integration
- Open-source codebase with community contribution guidelines

Future Extensions

Extension 1: Mathematical Content Processing

- LaTeX equation detection and MathML conversion
- Advanced equation accessibility with alternative representations
- Specific metrics for data integrity and mathematical markup robustness

Extension 2: Advanced Visual Content & Interactive Elements

- Scientific figure analysis and accessible chart generation
- Faculty + AI collaborative creation of visualizations, simulations, and interactive charts
- Data extraction from graphs and diagrams with accessible alternatives

Extension 3: Performance & Scale Optimization

- Long document processing with chunked rendering for documents >25 pages
- Cross-reference integrity maintenance across complex documents
- Performance optimization for institutional-scale deployment
- Advanced caching and CDN integration for high-volume processing

Extension 5: Production Automation

- Canvas Live Events webhook integration
- Automated processing workflows for existing PDF repositories
- Institutional reporting focused on accessibility compliance metrics

Payment Structure & Milestone Schedule

Deliverable-Based Payment Schedule ($60,000 Total)

Milestone 1: Project Initiation & Technical Foundation (25% - $15,000)

- **Payment Structure**: A project kick-off fee of **$7,500 (50% of Milestone 1)** is due upon contract signing to secure resources and begin architectural planning. The remaining **$7,500** will be invoiced upon completion of all Milestone 1 deliverables.
- **Key Deliverables**: Complete technical architecture, AWS infrastructure setup, multi-agent framework, basic PDF processing pipeline, project management setup
- **Acceptance Criteria**: Text extraction working, infrastructure deployed, progress report delivered

Milestone 2: Core System Implementation (30% - $18,000)

- **Key Deliverables**: All 6 agents operational, semantic caching system, web interface, first 10 test documents processed, confidence scoring implemented
- **Acceptance Criteria**: Complete multi-agent pipeline functional, UIC documents successfully processed

Milestone 3: Integration & Pilot Testing (25% - $15,000)

- **Key Deliverables**: All 30 pilot documents processed, Canvas integration demonstrated, faculty review interface, S3 hosting operational, accessibility compliance reports
- **Acceptance Criteria**: Full pilot scope completed, integration capabilities proven

Milestone 4: Final Delivery & Documentation (20% - $12,000)

- **Key Deliverables**: Technical documentation, open-source codebase published, training materials, performance benchmarking, Phase 2 roadmap, final project report
- **Acceptance Criteria**: Complete project handover with all documentation and future planning

Payment Terms

- **Invoicing**: Within 5 business days of milestone completion
- **Payment Timeline**: Net 30 days after invoice approval by UIC
- **Change Management**: Additional work requires new milestone agreement with separate pricing
- **Risk Mitigation**: Clear acceptance criteria for each milestone protect both parties

Budget & Resource Allocation

Operational Cost Projections

Pilot Infrastructure Costs (Foundation Phase):

- AWS ECS Fargate: $2-5/month (pay-per-use, ~6 hours processing time)
- Redis Cache: $3-8/month (minimal queue management usage)
- S3 Storage: $3-10/month (document hosting with versioning)
- CloudFront CDN: $2-5/month (content delivery for converted documents)
- **Total Monthly**: $10-28 for pilot usage (30 documents/month)

AI Processing Costs:

- **Cost per document**: ~$0.20 (typical 3-page document)
- **Cost variability**: Initial costs may fluctuate during testing and optimization phase; becomes highly predictable once processing patterns are established
- **Models**: Mix of cost-optimized (text) and premium (images/math) AI models

ROI Analysis:

- Traditional manual remediation: $15-36 per document
- AI-powered system: ~$0.20 per document + infrastructure
- **Cost reduction**: 95-99% vs. traditional methods

Institutional Value Creation

Immediate Benefits:

- Permanent accessible asset creation with institutional ownership
- Reduced dependency on external accessibility vendors
- Faculty empowerment through transparent AI partnership
- Compliance risk mitigation with proactive accessibility measures

Strategic Advantages:

- Academic publication opportunities on AI accessibility methodology
- Grant positioning with accessibility innovation requirements
- Open-source community leadership in academic accessibility
- Future-proof architecture independent of vendor dependencies

Strategic Value Creation:

- Institutional accessibility expertise and capacity building
- Community-driven open-source enhancement model
- Scalable infrastructure for comprehensive accessibility compliance
- Foundation for advanced AI-powered educational technologies
- Cost-predictable operation with transparent per-document pricing
- Vendor-independent architecture reducing long-term dependency risks

Risk Management & Limitation Acknowledgment

Technical Boundaries & Processing Limitations

Phase 1 Explicit Limitations:

The following will hold no guarantee on quality and will be flagged for degraded confidence until specialized dedicated work is put towards them.

- **Mathematical Content**: Complex LaTeX equations
- **Advanced Tables**: Merged cells and complex data relationships
- **Scientific Figures**: Complex accessible alternatives
- **Long Documents**: 50+ page optimization and cross-reference
- **OCR-Only Content**: Poor quality scanned documents

Processing Variability Expectations:

- **Success Rate**: 80% automation for common document types, 20% require manual intervention
- **Processing Time**: 1-20 minutes per document depending on complexity and size
- **Quality Assurance**: Faculty/TA oversight essential for institutional standards compliance
- **Edge Case Handling**: Clear error reporting with specific improvement recommendations

Security & Privacy Considerations

Details will be worked out with UIC IT and I. These are up-front core considerations I had in mind.

Data Protection:

- Course material processing only - no student education records or personally identifiable information
- PDF content analysis without personal data retention beyond processing duration
- Secure processing with automatic content deletion after conversion

System Security:

- AWS security best practices with encrypted storage and transmission
- Access control and authentication integration with UIC systems
- Backup and disaster recovery procedures for business continuity

Continuous Improvement Strategy

Quality Monitoring & Feedback Systems:

- **Equality Integration Tracking**: Before/after accessibility violation metrics and processing success rates
- **Faculty Correction Interface**: Simple UI for validating AI decisions and correcting semantic cache insights
- **Feedback Loop Infrastructure**: Redis-based system for storing corrections and improving future processing
- **Institutional Learning**: Corrections and validations contribute to UIC-specific processing improvements

Feedback Mechanism Implementation:

- **Foundation Phase**: Basic correction interface for headings, alt text, and confidence adjustments
- **Extension**: Enhanced correction workflows with semantic insight validation and reasoning chain editing
- **Infrastructure**: Simple web forms with API endpoints feeding back into semantic cache system

Conclusion

This comprehensive technical proposal outlines a sophisticated AI-powered system that transforms UIC's PDF accessibility challenges into institutional advantages. By combining transparent artificial intelligence with educator partnership, the system creates permanent accessible assets while building internal accessibility expertise.

The technical architecture leverages proven academic precedents and modern cloud infrastructure to deliver reliable, scalable accessibility solutions. The multi-agent processing approach ensures specialized handling of diverse document types while maintaining the transparency necessary for academic validation and community contribution.

Strategic Impact: UIC gains first-mover advantage in transparent AI accessibility, creating foundation for academic publication, grant opportunities, and community leadership while addressing immediate compliance requirements and student accessibility needs.

Technical Excellence: The system combines cutting-edge AI capabilities with proven accessibility standards, ensuring both innovation and reliability in institutional document processing.

Future Expansion: Clear technical pathway for advanced capabilities including mathematical content, complex visualizations, and institutional-scale automation, supported by demonstrated foundation system success.