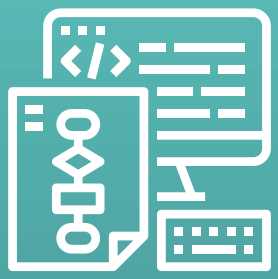


NLM Conversion to Build "Atomic" Physics Content in an Agile Fashion

A Case Study in High-Quality Legacy NLM Conversion

DCL™
Data Conversion Laboratory Inc.

OPTICA PUBLISHING GROUP
Formerly OSA

DCL
61-18 190th Street
Suite 205
Fresh Meadows, NY 11365
+1.800.321.2816
www.dclab.com

Co-authored by:
Optica Publishing Group
M. Scott Dineen, Sr. Director of Publishing Production & Technology
Alexander Schwarzman, Content Technology Architect

Data Conversion Laboratory
Mark Gross, President
Devorah Ashlem, Beth Friedman, and Gitty Kupferstein, Project Managers

ABOUT DCL

Intelligent
data transformations

DCL (www.dclab.com) provides data and content transformation services and solutions. Using the latest innovations in artificial intelligence, including machine learning and natural language processing, DCL helps businesses organize and structure data and content for modern technologies and platforms. With expertise across many industries including publishing, life sciences, government, manufacturing, technology and professional organizations, DCL uses its advanced technology and U.S.-based project management teams to solve the most complex conversion challenges securely, accurately and on time.

Your data:
transformed, validated, enriched

CONTENTS

NLM Conversion to Build "Atomic" Physics Content in an Agile Fashion

When faced with the challenge of converting eight highly technical journals spanning 95 years, how do you divide responsibility between the content owner and the conversion vendor? Do you spend a year on document analysis and developing conversion specifications, or do you hand the project over to a well-regarded service provider and rely on their expertise entirely? This paper demonstrates how an agile approach to content conversion with close collaboration between the publisher and the conversion vendor has allowed The Optical Society of America (OSA) and Data Conversion Laboratory, Inc. (DCL) to navigate between the two extremes and create a high-quality digital archive that will serve OSA's strategic aims for developing innovative products and services.

OSA is a scholarly publisher of 15 journals (some held in partnership with other societies) hosted on OSA's Optics InfoBase platform. As part of a five-year strategic investment to create more flexible products and services, OSA decided to convert its legacy journal content back to volume 1, issue 1, of The Journal of the Optical Society of America, which debuted in 1917.

Because older OSA journal content continues to be heavily read and cited—and because it contains items of historical interest to the society, such as biographies and meeting minutes—a decision was made to convert the entire journal back file to full-text XML.

The physicist leadership in OSA publishing wished to create uniform, well-polished "atoms" from the legacy materials for use in rapid development of new offerings, such as the Optics Image Bank and Enhanced HTML Article, mentioned below. When captured as XML, items such as equations, tables, and algorithms—which appear frequently in OSA journals—would be used for maximum product flexibility once freed from their PDF fetters.

In selecting an industry-standard conversion target, OSA chose the NLM Journal Publishing tag set, primarily because of its adoption by many of OSA's peer publishers and because of robust support by OSA's software providers and composition partners.

OSA would also have to make an important decision in selecting a vendor for the large conversion project. The high volume (more than 750,000 pages) and need for scientific accuracy (what happens when a single plus sign is dropped from an equation?) required a partner with both subject-matter expertise and a proven track record in handling large-volume conversion to an NLM XML target.

OSA chose Data Conversion Laboratory (DCL) following a thorough vetting process and found a partner that could work within OSA's budget constraints, bring significant experience with conversion to NLM

markup, and demonstrate robust and highly automated quality assurance.

Startup and Early Returns

OSA saw fairly quick return on its initial conversion investment. Within 18 months of engaging DCL, OSA had converted 6 years of its most-recent journal content and developed two well-received new offerings on its Optics InfoBase platform.

The Optics ImageBank (Figure 1), which provided a visual search and browse across all the figure images in available journal articles, gave researchers a way to locate and export scientific images using the context of the figure caption, the in-text reference, and related images as guides. The highly visual nature of optics science—with appealing subjects such as rainbows, holograms, and lasers—made an image bank a natural outgrowth product for the OSA community and one that OSA could not have developed effectively without reliable, uniform XML.

Along with the Image Bank, OSA debuted a full-text HTML version of its journal articles that offered conveniences such as tabular browsing of content, sortable citation lists, and reflowable text suitable for handheld devices.

Agile Approach

So this is a success story, but the focus will not be on OSA's vendor-selection process or its product-development efforts, which—truth be told—were the easier and more-enjoyable activities. The central challenge we faced, and still face as we pass the halfway point in the project, is to maintain a highly collaborative and agile approach to conversion specifications as we encounter one surprise after another in the legacy content. In spite of significant upfront planning and content inventory, the

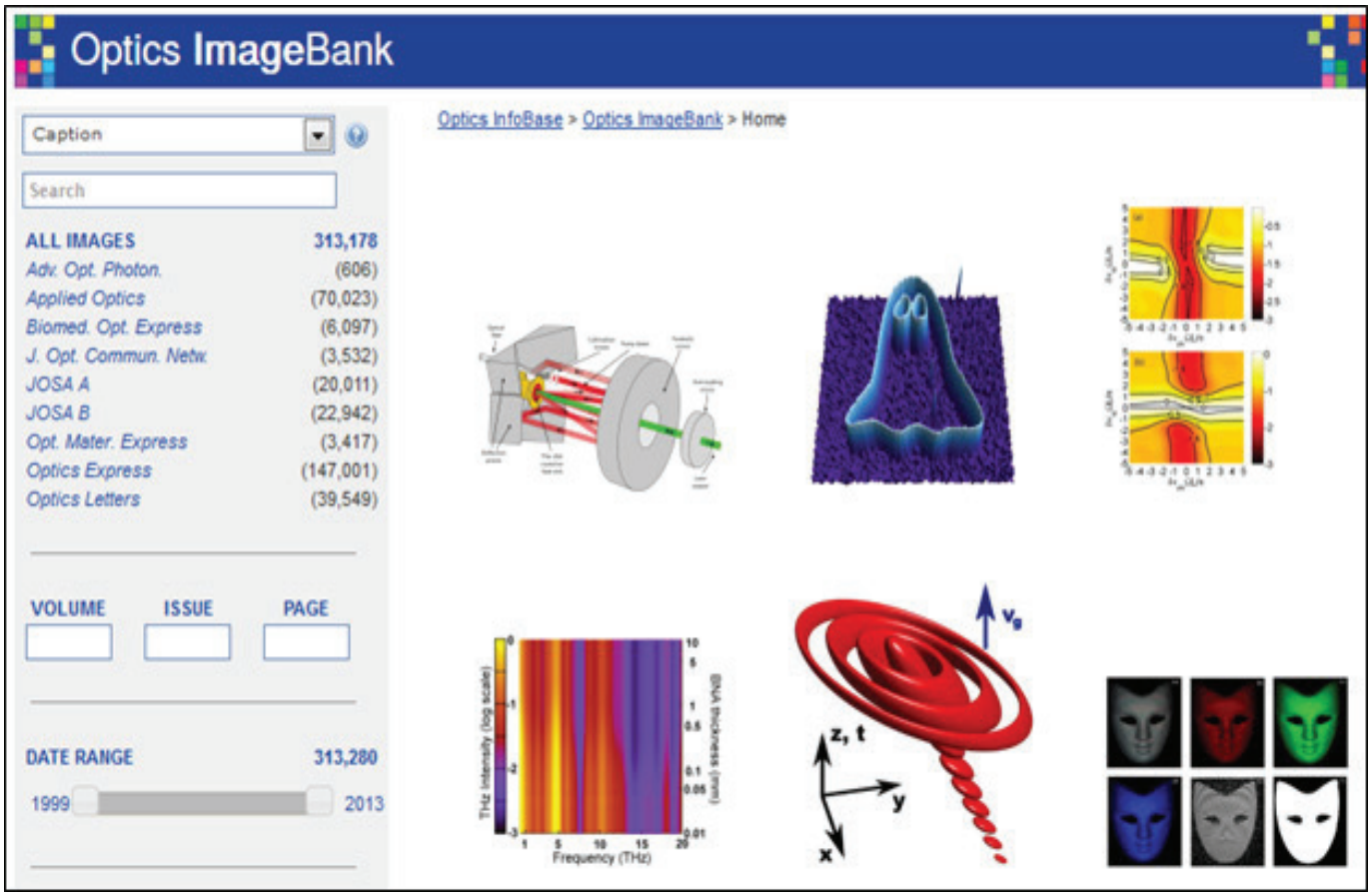


Figure 1: OSA's Optics ImageBank (imagebank.osa.org)

sheer number of articles and variety of legacy input formats made traditional comprehensive spec writing impractical.

The balance of the paper will illustrate how project responsibility has been allocated between OSA and DCL and how the most-problematic content-capture issues have been addressed in a collaborative, agile fashion within the constraints of the NLM Journal Publishing tag set.

Conversion as an Agile Process

OSA and DCL decided that performing a full inventory of all material and building detailed specifications for each situation would take months, add significant cost, and ultimately be deficient. Methodical sampling seemed a fair compromise with the understanding that continual reassessment would be needed. During project startup, we examined random, stratified samples to develop initial specifications. One aim was to ensure that the samples were indeed representative.

The "agile" process we pursued did not mean "no planning"; it was planning with the expectation that things would change. The project commenced with the practice of daily e-mail communications and regular weekly meetings to discuss handling each problem as it arose to ensure that outcomes were both practical and well-aligned with OSA's business goals.

Developing a Phased Approach

Any large legacy conversion effort can be daunting. There are many unknowns, and since it's usually impractical to fully inventory the collection, it's just not possible to plan everything in advance. As mentioned, OSA's journal corpus incorporated a large collection of complex materials, with many journal titles, each with its own personality, and many source types—PDF, XML, and SGML—each with its own nuances. Further burdening the effort was the long life of some titles, spanning almost 100 years, with the attendant changes over time in style and substance.

Figure 2: DCL Conversion Hub. What to Use as the Source (or I Thought Converting from SGML Was Easy)

The overall plan was to organize the project into phases that allowed production to start early on the better-defined materials; this approach would allow parts of the collection to get online soon while at the same time improving workflow effectiveness over the project's 4-year duration. OSA and DCL worked closely to define the phases based on technical and business considerations.

Some materials were already tagged to OSA's own XML DTD, which was highly mapable to the NLM 3.0 DTD. This content would be the quickest to move over and would result in 6 years' worth of content being quickly available. It also allowed for initial new products to be developed by OSA in short order. From there DCL moved on to older content in other XML formats, and then to PDF and SGML formats.

The approach was to work in phases by source type, and within source type by title, which allowed us to better focus on a title's nuances. Within a title we would start with the most current materials and work backwards.

Building Flexibility into the Process

Given the wide range of materials and the number of unknowns, we knew we would need to build flexibility into the conversion process. We developed a formal specification as a starting point, with the understanding that it would be continually revised as new data situations arose. The software was developed in sprints for incorporating new scenarios as they developed. The downside to this approach was that sometimes we discovered changes that affected previously converted materials. The close working relationship we had developed allowed us to arrive at mutually acceptable compromises on how to manage these changes.

These important tools were used to retain flexibility over the course of the project:

1. Hub-and-Spoke Infrastructure. DCL's software process is based on a hub-and-spoke infrastructure (Figure 2), which allows for multiple source types to be taken in and pre-processed appropriately to their source type, and then further processed through a common set of data modules. This hub-and-spoke approach, part of DCL's toolset, is specialized to handle front matter, special characters, several varieties of tables, equations, and other structures. Modules are added as needed to handle newly discovered structures. That all sources can take advantage of the same toolset allows greater consistencies to be achieved over a large corpus of material.

2. Quality-Assurance Software. Automated tools to perform quality assurance (QA) on the converted files also provide valuable information on new structures that start appearing as new materials are processed, and in particular as you go further back in time in a legacy collection, finding structures that don't follow the initial rules. It's effective to have a process that allows one to track these changes methodically over time.

3. Learning Databanks. The changes over time can be used to continually build databanks that facilitate ever-improving QA. An example shown later in this paper is the hyphenation checker, which can be used to differentiate between hard hyphens, which should be retained, and soft hyphens, which should be removed when converting to XML. Over time further examples of correct hyphenation can be collected and used on newly converted materials.

While one would expect that converting from highly structured formats like SGML and XML would be the easiest, that's not always true, as we'll discuss below. Each source presents its own unique difficulties, and some of these are counterintuitive.

Converting from XML. The first phase consisted of converting a recent collection that had been tagged in XML using OSA's proprietary DTD. These recent materials used structures similar to those used in NLM

3.0 and had been created very consistently and accurately. As a result we were able to map most structures directly to those needed in the target NLM 3.0 DTD and were able to create a highly automated conversion. DCL and OSA worked together to incorporate the OSA-provided rules into DCL's conversion software, which was able to accurately clean up and normalize content in the course of conversion. Another XML-based collection was not as easy to convert. These materials, already in NLM 2.3, needed just to be upgraded to NLM 3.0. On its face, this should have been easy and highly automated. However, in this case the source files had not been created very carefully and had significant flaws. As a result there was a need for additional software-based QA, along with extensive visual review for a number of problem areas—such as the converted MathML.

Converting from PDF. We next tackled materials that had been sourced in PDF.

- PDF Normal. Some of OSA's materials were in a text-based PDF, which contains extractable text. While these don't usually need proofreading, there are a number of problem areas that need to be addressed, since these files are designed for printing and not for conversion. Areas that need review include hyphens and special characters, and there's the general problem that the XML structures you need are not in the PDF and have to be reverse-engineered.

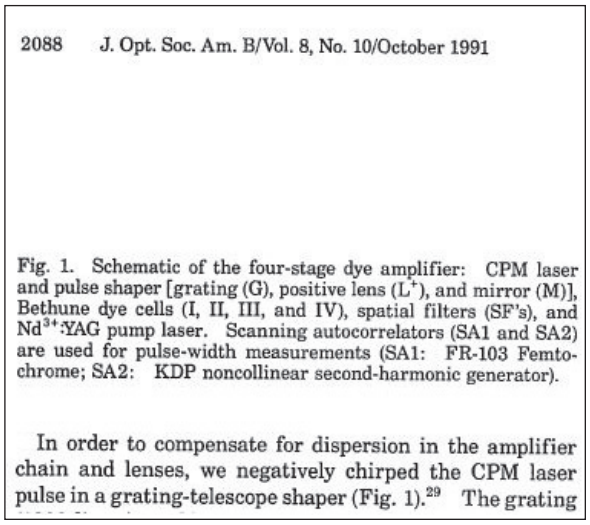
- PDF Image. The majority of OSA's collection was scanned image PDFs. These had to be OCR'd and proofread before they could be converted. While the conversion of proofread content was highly automated, there was an extensive QA process to ensure accuracy of conversion, especially for older materials with unusual structures.

Converting from SGML. SGML is highly structured like XML, retains full text, and should theoretically be simple to convert. However, SGML as source presents its own set of challenges. These were older materials, and a big hurdle was identifying the authority versions of documents. In addition there was the wide variety of SGML input that had been produced by various vendors over time. Coming from an era before much standardization, there were many variants of tagging, and coming from a variety of vendors, there was inconsistency in the quality of the source.

Another challenge in converting directly from SGML source is that some tagging that is valid in SGML does not work properly when converted to XML. For example, in SGML, it is valid to omit table columns since they have IDs, but when mapped to XML with the columns omitted, the table appears skewed. To produce good-looking tables, the conversion software needs to analyze and account for the "missing" cells, adding spanning where necessary, thus completing the table.

Because of the wide variations found, OSA and DCL worked collaboratively to reassess feasibility of the automated approach from SGML, as opposed to working from the PDF. In reviewing results together, we were able to mutually agree on what would be acceptable output from the SGML conversion. Only tagging changes that wouldn't compromise the quality of the output were allowed. For example, the source SGML had long equations tagged as one long line—which was different from the corresponding PDFs. Since MathJax (a browser-based display engine for math) has the capability of breaking the equations on the rendering end, the resulting XML was still deemed accurate. We are currently working through the SGML conversion on a title-by-title basis as we've done for the other source materials. As we convert new titles, we plan to reassess anew the quality and consistency of each SGML set.

Other Complexities. Aside from multiple sources, there are other complexities that required collaborations between OSA and



2088 J. Opt. Soc. Am. B/Vol. 8, No. 10/October 1991

Fig. 1. Schematic of the four-stage dye amplifier: CPM laser and pulse shaper [grating (G), positive lens (L^+), and mirror (M)], Bethune dye cells (I, II, III, and IV), spatial filters (SF's), and Nd^{3+}:YAG pump laser. Scanning autocorrelators (SA1 and SA2) are used for pulse-width measurements (SA1: FR-103 Femtochrome; SA2: KDP noncollinear second-harmonic generator).

In order to compensate for dispersion in the amplifier chain and lenses, we negatively chirped the CPM laser pulse in a grating-telescope shaper (Fig. 1).[29] The grating

Figure 3: Article with missing graphic

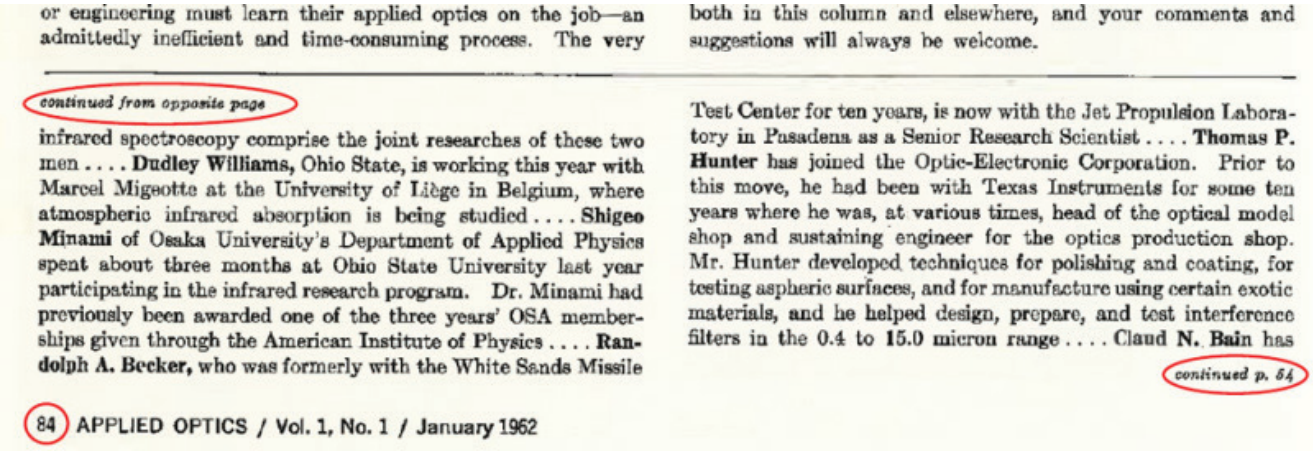
DCL. Aside from the actual conversion of content from one form to another, a goal of the conversion process was to normalize data to make it fit the new structures in a consistent manner. As mentioned, some structures didn't fit neatly within the constraints of NLM 3.0, and we had regular meetings to discuss how best to remap these structures appropriately. Some examples include the CALS to HTML table conversion, MathML line break retention, and explicit tagging.

Then there were the content issues. With the vast amount of content, we found a number of unexpected content scenarios and inconsistencies that required collaborative decision-making. Missing text, pages that jumped from one part of the journal to another, and character table alignment are a few examples.

Some of the content issues needed to be dealt with on a case-by-case basis. For example, we did find errors of omission in some of the older material and did come up with general rules, which we reviewed periodically. For a missing graphic in a figure that was actually referenced in the article (Figure 3), we would want to scan the blank space as a graphic placeholder, retain the figure caption, and assign an ID so that it could be referenced in the article.

The text [figure omitted on the printed page] would be inserted into the caption so that the user would know that a graphic is missing. In other cases, if it was just plain text that was omitted, we could just include a placeholder such as [missing text] inserted into the paragraph text.

In some instances we found that pages jumped from place to place within the journal, sometimes jumping backwards (Figure 4). To accurately represent these types of occurrences, OSA decided on a trade-off. Rather than have all of the text converted into XML, DCL would spend additional effort to combine all of the scattered pages of an article into a unified PDF and then just tag the header information. That way, the information would



or engineering must learn their applied optics on the job—an admittedly inefficient and time-consuming process. The very

both in this column and elsewhere, and your comments and suggestions will always be welcome.

continued from opposite page

infrared spectroscopy comprise the joint researches of these two men . . . . **Dudley Williams,** Ohio State, is working this year with Marcel Migeotte at the University of Liège in Belgium, where atmospheric infrared absorption is being studied . . . . **Shigeo Minami** of Osaka University's Department of Applied Physics spent about three months at Ohio State University last year participating in the infrared research program. Dr. Minami had previously been awarded one of the three years' OSA memberships given through the American Institute of Physics . . . . **Randolph A. Becker,** who was formerly with the White Sands Missile Test Center for ten years, is now with the Jet Propulsion Laboratory in Pasadena as a Senior Research Scientist . . . . **Thomas P. Hunter** has joined the Optic-Electronic Corporation. Prior to this move, he had been with Texas Instruments for some ten years where he was, at various times, head of the optical model shop and sustaining engineer for the optics production shop. Mr. Hunter developed techniques for polishing and coating, for testing aspheric surfaces, and for manufacture using certain exotic materials, and he helped design, prepare, and test interference filters in the 0.4 to 15.0 micron range . . . . Claud **N. Bain** has

continued p. 84

84 APPLIED OPTICS / Vol. 1, No. 1 / January 1962

Figure 4: Article with "jumping" pages

Table I. Sin

	No.	1
$f_{(film)}$	$f_{(Eye)}$	Eye
2.5	9.2	0.93
7.5	27.6	0.60
12.5	46.	0.31
17.5	64.5	0.18
22.5	84	0.08
27.5	101	0.03
32.5	119	0.01
37.5	138	

Figure 5: Source PDF table with decimal alignment

be easily accessible and readable in one flowing PDF.

Other conversion issues related to layout in the source that isn't supported in either the NLM 3.0 DTD or its rendering. For example, decimal alignment isn't available in most HTML rendering, so we wanted to avoid it. Therefore, for source tables using decimal alignment (Figure 5), the choices were either right, left, or center, and so we developed rules on what to do. After reviewing the options, we decided to center the decimal-aligned columns, as it displayed the information most legibly.

In other cases, the source document used special characters with no corresponding Unicode equivalent entities, for example, chemical bonding symbols (Figure 6). In this case, we needed to find the right balance and compromise to make sure the material would render as well as possible—while staying accurate to the source. Our approach was to work collaboratively to review the options and settle on a course of action.

Quality Assurance

As described above, quality assurance (QA) is used both to identify errors and to provide feedback for process improvement. This was another area of extensive collaboration between OSA, with its extensive subject-matter expertise, and DCL, with its expertise in the conversion process.

The QA process as used in this project was multiphased and included visual reviews and automated steps:

1. Visual QA
2. Schematron
3. Reporting stylesheets
4. OCR/hyphen/spelling checking software
5. Quality-control software

Visual QA is performed on each article assisted by several DCL viewing tools, which provide visual indicators for identifying errors. Likely trouble spots are highlighted using colors, fonts, and point sizes (Figure 6) that make it easier to identify elements with a quick look. Items highlighted include Unicode characters, alignment, spanning, and borders—all items which are easy to overlook even by the most careful reviewer.

Visual review of math is done with MathJax and is used by both OSA and DCL. It is very helpful that we are both using the same viewing tools for math, since the math is captured as code that must subsequently be rendered (and there are variations in

Characterization of the polymerization rate constant in an acrylamide-based photopolymer for holographic recording using Raman spectroscopy has been presented. The consumption of monomer was observed to be monoexponential. A time constant from the exponential fit of the intensity peaks corresponding to an acrylamide carbon–carbon double bond (C═C), a carbon–hydrogen vinyl bond (CH_2), and a carbon–carbon double bond of bisacrylamide (C═C) was obtained and the polymerization rate constant was determined. It was determined experimentally that the dependence of the polymerization rate on the

Figure 6: QA visual rendering. Box-drawing entities used to represent chemical bonds

Source	Article-Id	Vol Issue	Pages	DOI	Pub-Date Info (month year)
josab-10-10-1801.xml	josab-10-10-1801	10/10	1801-1809	10.1364/JOSAB.10.001801	EPUB: D01 M10 Y1993 COLLECTION: D01 M10 Y1993
josab-10-10-1810.xml	josab-10-10-1810	10/10	1810-1819	10.1364/JOSAB.10.001810	EPUB: D01 M10 Y1993 COLLECTION: D01 M10 Y1993
josab-10-10-1820.xml	josab-10-10-1820	10/10	1820-1823	10.1364/JOSAB.10.001820	EPUB: D01 M10 Y1993 COLLECTION: D01 M10 Y1993
josab-10-10-1824.xml	josab-10-10-1824	10/10	1824-1833	10.1364/JOSAB.10.001824	EPUB: D01 M10 Y1993 COLLECTION: D01 M10 Y1993
josab-10-10-1834.xml	josab-10-10-1834	10/10	1834-1839	10.1364/JOSAB.10.001834	EPUB: D01 M10 Y1993 COLLECTION: D01 M10 Y1993

Figure 7: QA stylesheet report

how different software supports and renders MathML source code). MathJax lends itself well to the presentation of the math captured in the OSA articles, as it can render long equations tagged as one line on multiple lines, making the math notation easier to read and understand.

Schematron is a key tool for identifying many XML issues. The extensive Schematron libraries we use were developed by OSA initially as a way to verify that deliveries followed the OSA guidelines that we had mutually developed, but ultimately we found Schematron valuable for both OSA and DCL to use to keep track of the current rules. The Schematron code is still updated on a constant basis, as specs are changed, so we have an up-to-date check of what the tagging should be. Schematron checks many tagging issues such as article type, bibliographic accuracy, etc., and is run against all files as part of the QA process.

Other software checks are used to test for conditions that would be difficult for Schematron. XSLT stylesheets are used to create reports on chunks of data that need to be broken into components. The report pictured in Figure 7 displays a column for each tag, making it easier for an editor to review the contents. For example, to see how the metadata is tagged, a stylesheet is run against the content that creates a report. This is visually easy to review and find errors. This same approach can be used for references, author names, and any other elements for which a breakdown of the text is necessary and needs to be checked to inspect content capture.

When the source material is image PDF, an additional level of quality checking is required for ensuring that all the text was extracted and OCR'd properly. To help with the OCR checking, we created modified versions of the fonts designed to help distinguish between similar-looking characters (e.g., "O" vs. "0", "Z" vs. "2", "1" vs. "l") used within the proofreading phase (Figure 8).

Another tool that DCL uses, particularly with PDF sources, is a hyphenation spellchecker (Figure 9), which catches extraneous soft hyphens or missing hard hyphens. For words that can be spelled with or without a hyphen (for example, well-known), the software checks for frequencies of words appearing with and without hyphens, to determine which words should have their hyphens dropped.

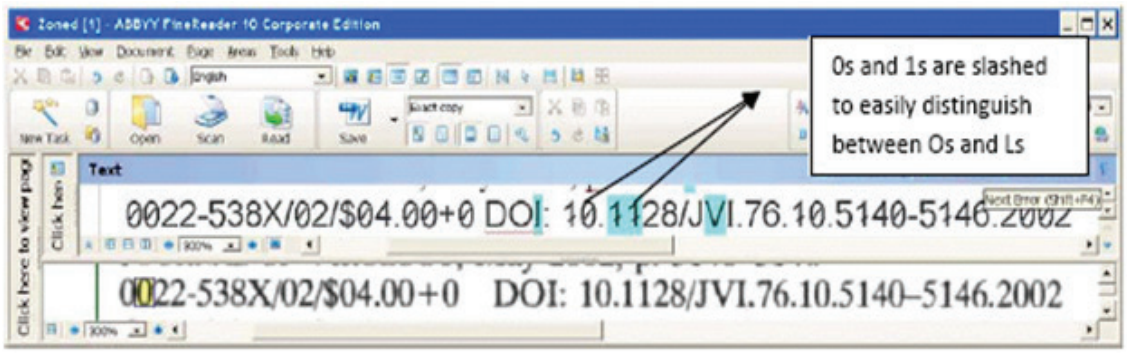


Figure 8: QA review font

Word with Hyphen Variants	
phospho-lipids	1
Phospholipids	2
phospholipids	32
transcrip-tome	1
transcriptome	2
metabolo-mics	1
metabolomics	1
Metabolomics	3
lipo-proteins	1
Lipoproteins	3
lipoproteins	23
retinal-dehyde	1
Retinaldehyde	3
retinaldehyde	61
knock-out	1
knockout	4
glycosphingo-lipids	1
glycosphingolipids	3
P-450	2
P450	11
up-regulate	1
upregulate	1
neuroblas-toma	1
neuroblastoma	1
mu-rine	1
Murine	1
murine	7

Figure 9: QA hyphenation report

Other post-conversion software-based quality checks help to identify discrepancies between the XML files and the tagging specifications, and checks for global content issues such as ensuring file completeness, ensuring all external files are accounted for, and identifying paragraphs that begin suspiciously (Figure 10).

Quality assurance is critically important in any conversion project, but the challenges with legacy scientific articles are formidable and lapses unforgivable, as even the smallest omissions or modifications can result in misrepresenting the science.

Conclusions

The OSA is making a sizable, multiyear investment in full-text XML conversion of its journal back file to allow for nimble reuse of content. The legacy journal content holds historical and scientific value, and the OSA has already used a portion of the converted material to build engaging new interfaces for image discovery and HTML article display. The NLM XML will form the basis for additional enrichment around topical areas, author profiles, and other meaningful items within the articles. Without the granularity, uniformity, and quality of the tagging, OSA would have far fewer options for proceeding confidently in development of new offerings.

In outsourcing a task as large and complex as this, it would be tempting for the publisher to identify a well-regarded service provider, hand responsibility over to the experts, and be done with it. Why would a publisher want to outsource a conversion project to recognized experts and also tie up senior internal staff over a four-year project cycle?

Consider the following summary points:

- Flexibility to change direction quickly is critical, but in many cases only the customer can ensure that the change aligns with business needs.
- Monitoring quality from both sides is important. The client and vendor are bound to have variations in QA environments that reveal problems in different ways.
- With the right partners, the collaborative environment improves morale, attention to detail, and decision-making.

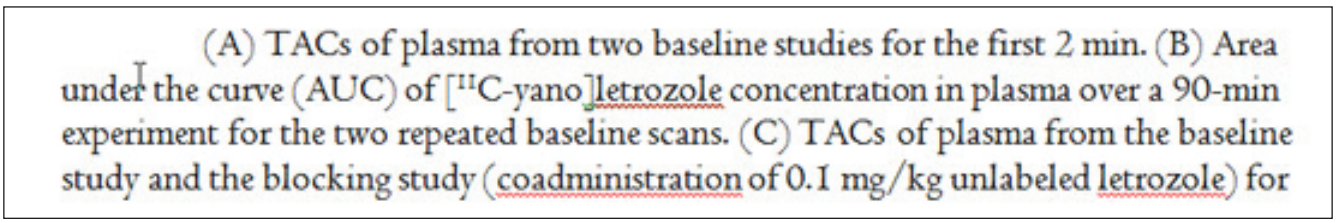

(A) TACs of plasma from two baseline studies for the first 2 min. (B) Area under the curve (AUC) of [^{11}C-yano]letrozole concentration in plasma over a 90-min experiment for the two repeated baseline scans. (C) TACs of plasma from the baseline study and the blocking study (coadministration of 0.1 mg/kg unlabeled letrozole) for

Fig. 10: Suspicious start of paragraph

DCL™

Data Conversion Laboratory Inc.

Address	61-18 190th Street Suite 205 Fresh Meadows, NY 11365
Telephone	+1 800.321.2816
Web	www.dclab.com
Twitter	@dclaboratory
LinkedIn	linkedin.com/company/dclab