Content Structure: The Building Blocks of Innovation

Extracting & Enriching Data to Accelerate Digital Transformation

a report from

Data Conversion Laboratory
61-18 190th Street
Suite 205
Fresh Meadows, NY 11365
+1.800.321.2816
www.dataconversionlaboratory.com
ABOUT
Data Conversion Laboratory

Intelligent Data Transformations

Data Conversion Laboratory 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.

Your data: transformed, validated, enriched
CONTENTS

04 FOUNDATION FOR SUCCESS

04 FROM FLAT TO MULTIDIMENSIONAL

06 TECHNOLOGIES OF TRANSFORMATION

06 THE ROLE OF AUTOMATION

07 THE POWER OF INTELLIGENT CONTENT

08 DATA QUALITY IS KEY

08 THE DCL DIFFERENCE

09 TAKEAWAYS FOR YOUR ORGANIZATION

Organizations Interviewed for this Report

05 THE NEW YORK PUBLIC LIBRARY

07 THE AMERICAN WATER WORKS ASSOCIATION
Content Structure: The Building Blocks of Innovation

Extracting and Enriching Data to Accelerate Digital Transformation

Foundation for Success
All organizations have deep, growing archives of content. From historical documents and photographs to new research articles and industry standards, this content comes in many different forms, yet it can all be mined to create valuable resources for today’s digital-first consumers.

However, content can only unlock new opportunities for organizations if it has a foundation of rich structure and metadata. Simply having information in a digital format, such as a Word file or an image-based PDF, is not enough. For content to be easily discovered and used via modern platforms, it must be converted into a multidimensional XML format from which machines can extract pertinent information.

Thankfully, artificial intelligence technologies, including natural language processing and machine learning, have lowered previous barriers to constructing rich XML. Content-transformation projects require less time and fewer human resources than ever. In partnership with a technology solutions provider, content-producing organizations can undertake once unfeasible projects – in turn, making their wealth of resources more accessible and monetizable through digital product offerings.

This report explores how any organization can leverage new technologies to create intelligent, multidimensional content. By extracting and enriching data, organizations accelerate digital transformation to innovate, meet modern consumer expectations, and propel business forward.

From Flat to Multidimensional
Before embarking on a data conversion and content structure project, organizations must take stock of the information they have buried within their files, and determine what is worth enriching and making digitally available. Likely, information is stored in a variety of formats, including hard copies (e.g., papers and microfilms) and digital copies (e.g., PDFs and word processing files).

“An organization that has been collecting information for many years just doesn’t know what they have,” says Mark Gross, President and CEO of Data Conversion Laboratory (DCL). “Everybody is so busy doing their day-to-day jobs that this is just one of those things that gets pushed to the side, even though it could be a very lucrative source of income for the organization. One of the things DCL can do for an organization is just an inventory of what they have - and often what they have is a combination of different kinds of things.”

For example, DCL helped The Optical Society of America (OSA) inventory its materials in
preparation for a 100th anniversary initiative to make scientific content from the OSA archives available online in a searchable repository. OSA’s content existed as paper, microfilm, and outdated XML that needed to be standardized, so DCL reviewed the mix of materials and determined the right tools and approach for each particular source of information.

Oftentimes, organizations have digitized information, but that information is locked in image-based PDFs and not structured to facilitate search on an updated platform. To unlock new dimensions of discoverability, the information within those images needs to be identified and tagged in a way that is conducive to search and utility (not to mention making it accessible for the visually impaired).

This was the case for The New York Public Library (NYPL), which had scans of historical records from the United States Copyright Office that it wanted to enter into a searchable database to quickly assess the copyright status of books and other registered works. NYPL partnered with DCL to extract and parse data from the scanned images. Beginning with a 10,000-page pilot set of records published between the years of 1923-1964, the United States Copyright records were transformed into XML.

The first step was to transcribe the data embedded in the scans. “For some materials, using automated OCR (optical character recognition) works well enough,” says Gross, “but not when you’re dealing with complex materials like the copyright records, which contain multiple columns and components that are hard for a computer to distinguish from one another.” For this reason, DCL staff developed custom software to scan and transcribe data from the PDF files.

Before the data could be parsed into XML and loaded into a database, NYPL content experts worked closely with the DCL team to standardize a set of fields for markup that would enable accurate tagging of each component in the copyright record entries, notes Sean Redmond, Senior Product Manager at NYPL.

Figure 1. Content structure and technology enables interactivity and discovery from content that was previously static and “flat”.

© 2020 DCL     info@dclab.com    +1 718.357.8700
“We don’t just want to extract the text,” Redmond explains. “We want to know what does the text mean, beyond just the simple letters that are on the page. What part is the title? What part is the author’s name? What part is the registration number?”

Due to the complicated format of the source material, as well as inconsistencies in record-keeping through the years, the field standardization process required a unique combination of human resources and technology.

“For example, lots of publishers are people. For a computer to recognize that this first name and last name is an author, and then two lines down, this first name and last name is a publisher… you can only know that because there are human beings who recognize that name, because they’ve seen it before,” Redmond explains. “When people figure that out, they can put it into something that can be used over and over again in a computer.”

Technologies of Transformation

Artificial intelligence (AI) and natural language processing (NLP) are transforming the data conversion and structure process and opening previously closed doors for organizations. These technologies make it possible to automate accurate content markup, which allows businesses to revisit high-value but formerly impractical projects.

While AI is not a new concept, it requires a lot of computer processing, which prevented most organizations from leveraging AI capabilities in decades past. In the current technological landscape, however, computers are faster and more available, so organizations can now take advantage of AI to restructure vast amounts of content very quickly and with fewer resources. NLP, a subset of AI, is becoming particularly instrumental in content-transformation projects because it enables computers to better decipher and process text and human language.

The Role of Automation

“NLP gives us the capability that we didn’t have before to automate a process that formerly had to be done by people,” says Gross. “It’s reduced the cost of doing a lot of this kind of work to the point where it’s practical now when it wasn’t practical 10 or 20 years ago. If a project was looked at 10 years ago and wasn’t deemed practical, then it should be looked at again because the economics of it and the parameters have totally changed over the last few years.”

Automation isn’t just a gateway for new data-conversion projects; it also allows organizations to enrich outdated XML for previously created resources, ensuring functionality via modern interfaces.

“Once you digitize things, you have to keep maintaining them,” Redmond stresses. “You have to keep at least paying for the electricity for the computer that’s storing the information and keeping a search interface that’s accessible to the public. [Digital resources] take constant updating as computers change, browsers change, and iPhones are invented.”
In one such maintenance scenario, global information analytics business Elsevier enlisted DCL’s help to standardize tagging of entries in its Scopus citation database, a bibliographic guide to scientific literature that contains millions of entries. While entries from recent years had robust tagging, older content in the database did not.

“Elsevier wanted to go back and redo the preceding years so that everything had all those tags in it, but they wanted to make sure everything was done completely automatically,” says Gross. “DCL built software that we ran for them, and it went against another 150 million entries and added the tags to the right places using NLP.”

Another notable use of AI is for conversion of content that requires security. By automating processes, organizations can pass sensitive materials through fewer human hands — if any. For example, the U.S. Patent and Trademark Office partners with DCL to automatically transform confidential, complex forms and image-based PDFs to XML via a secure system, which ensures proprietary information is kept private.

AI technologies continue to evolve and improve, and the organizations that adopt them to make information more accessible to modern users will have a competitive advantage.

“The bar is really rising for how people look for information,” notes Gross. “There’s too much content out there, and more than anyone can get a hold of. What’s important today is how to find that particular piece of information that you need. In order to do that, you need to identify what’s in your content and pull out the metadata in a way to search for it accurately. Organizations who don’t do that will be left behind because the others will have it.”

The Power of Intelligent Content

With rich data structure comes a world of new content-powered possibilities. As Dan Berger, Senior Manager of Production at the American Water Works Association (AWWA), puts it, “The value is what you do with the XML. Converting to XML just means you can get there.”

A nonprofit scientific and educational association for water industry professionals, AWWA publishes international standards and best practices related to the treatment and management of water. AWWA worked with DCL to convert its technical standards into a specific XML format — the National Information Standards Organization’s Standard Tags Suite (NISO STS). AWWA’s standards and manuals were previously available in print and downloadable PDFs, and the association wanted to make these materials more searchable and readily available to the engineers and other industry professionals who need to use them. In just 8 months, DCL converted more than 22,000 pages of AWWA content.

“The point of this was to create a better experience for our customers, to create a new product for them,” says Berger. “Having things in a digital format allows us to put them onto a digital platform, and then that opens up all kinds of opportunities that are just unavailable when you’re dealing with a flat PDF.”
Intelligent content structure improves indexing and discoverability of AWWA materials across multiple channels — including websites, apps, and mobile — which empowers users to swiftly identify the right pieces of information. It also allows AWWA to include new, more interactive media within its standards, and will enable experimentation with augmented reality features down the road, says Berger.

For NYPL, a searchable copyright record database has improved access to literary and artistic works for the public and other informational institutions. NYPL can now easily pinpoint books that are in copyright but out of print, and then come to an agreement with rights holders to make materials available digitally. Using the database, NYPL also learned that a large portion of books published before 1964 are actually in the public domain, due to changes in copyright law and renewal contracts over time.

“You can do something with one purpose in mind, but the great thing about having the digital records is that you then start discovering new things. That’s what we’ve found with this project,” says Redmond. “Being able to search and manipulate and play with the information through a computer makes a whole lot of other things possible.”

By converting and combining its expansive archives into one database, OSA was able to build popular new products for its customers, including special collections on famous physicists and a searchable repository of vintage photographs that were previously buried. “The organization used what it already owned to create new sources of income,” Gross emphasizes.

For modern publishers and information providers, structured data is an essential building block for innovative new business models.

**Data Quality is Key**

In order to harness the power of intelligent content, organizations must go beyond simple data conversions. Not all XML is created equal, and extracted metadata needs to be high-quality for it to support continued business evolution.

Gross describes the difference between outdated, “lazy” XML and rich, intelligent XML using the example of a bibliographic entry in a book or article. The entry contains dozens of informational components (e.g., author’s name, article title, publisher, journal title) that should get tagged in the data conversion process.

“It could be 20 different pieces of information in that one little paragraph, and XML lets you tag each of those components to tell us what it is,” Gross explains. “But if you were lazy, you could just say, ‘the following is a bibliographic entry and everything else is just text.’ If all you did was identify the paragraph as a bibliographic entry, and now you’re moving to a system where you want to be able to identify author’s names, now you have to go in there and retag it.”

With proper due diligence and tagging, organizations set themselves up for success not only in their current endeavors, but also for future endeavors yet to be identified.

As Redmond advises, “You have to think about not just ‘how am I going to use these records today,’ but what information is really in these documents that should be encoded so that when we’ve done our first intended project with it, we still have more value we can get out of the process that we didn’t know we were going to want when we started.”

**The DCL Difference**

DCL differs from many service providers because it uses a combination of onshore staff plus technology to create well-structured data for its customers. The company’s deep knowledge of markup languages enables it to use leading technologies, such as NLP and ML, to generate accurate markup in
an automated, cost-effective way. DCL’s experience with standards XML, in particular, stood out to AWWA when it was searching for a technology firm.

“There are lots of kinds of content that can be converted into XML and lots of companies that can do it, but we were looking for somebody who has experience converting standards, which have their own way in which they get created,” says Berger. “We also wanted someone who had familiarity and was able to adapt to a new standard of XML (NISO STS) that was our core standard.” NISO STS was established in late 2017, and when AWWA partnered with DCL in 2018 it became one of the first standards for organizations to convert to the new “flavor” of XML.

DCL’s transformation services and markup language expertise benefit a wide range of industries. The publishers, libraries, and associations referenced in this paper are just a small subset — DCL also works extensively with organizations in financial services, manufacturing, government, legal, and many other markets, each with specific requirements for content structure. As such, DCL prioritizes continued research and technological innovation to help each client extract maximum value from their efforts.

No matter what new technologies bring, however, one thing is certain: “If your content isn’t tagged properly, people won’t find your information,” says Gross, “and making information more findable is critical for the success of organizations moving forward.”

Takeaways for Your Organization

• Today’s consumers have higher expectations for search, and information-providing organizations must restructure content to meet their needs.

• Artificial intelligence has lowered the barrier for embarking on data-conversion projects.

• Intelligent content enables product innovation, unlocking new lines of revenue for businesses.

• Well-structured data is paramount to true digital transformation.
Address  
61-18 190th Street  
Suite 205  
Fresh Meadows, NY 11365

Telephone  
+1 800.321.2816

Web  
www.dataconversionlaboratory.com

Twitter  
@dclaboratory

LinkedIn  
linkedin.com/company/dclab