

Document Imaging Report

Business Trends on Converting Paper Processes to Electronic Format

4003 Wood Street ● Erie, PA 16509 ● PH (814) 866-2247 ● <http://www.documentimagingreport.com>

July 22, 2016

Leveraging Crowdsourcing to Complement Capture

We spend a lot of time in *DIR* discussing automated recognition technology and how it can be used to improve the document capture process. The idea is that utilizing computers and software to extract information from document images is more efficient than having people key it. The reality is that this is not always the case.

I remember either Chris Thompson of RRI or Scott Blau of Datacap (one of these industry forefathers) explaining to me early on in my career, “Our biggest competition is not the other vendors in this market, it’s the people doing offshore key entry.”

And while automated capture software has certainly evolved over the years, so has outsourced key entry. The term “offshore,” is no longer in use—it originated at a time when paper documents were literally shipped to foreign locations where keying was done. With the growth in adoption of broadband Internet came the ability to transfer large quantities of document images online for key entry in countries like India—which developed a strong document capture outsourcing presence. More recently, we’ve seen the emergence of the cloud give rise to a crowdsourcing movement, which introduces almost unlimited resources to the key entry mix.

DIR recently caught up with a start-up that is attempting to leverage the rise of crowdsourcing and combine it with today’s automated recognition technology to create extremely accurate and cost effective data capture services for its customers. “If you look at most

automated capture operations in place today, they get between 60-80% of fields extracted accurately,” said Olaf Malchow, CEO of **ScaleHub**. “Our proposition is that we can hook into their capture operations and have the 20-40% of exceptions that need to be looked at by a person submitted to our crowdsourcing operation.

Then, when we give our customers back their data, it will be 100% accurate or close to it.”

ScaleHub, which is based in Germany, launched in February. So, technically it is a start-up. That said, its principals have been active in the data capture market since the late 1990s.

Malchow and COO Ralf Göbel were co-founders of Foxray, an ISV that developed a high-volume capture workflow platform. CTO Dan Dubiner spent several years at **Top Image Systems** before co-founding FasterAP, which is where he started leveraging crowdsourcing for data entry.

“Foxray was acquired by ReadSoft in 2012, and in 2014 **Lexmark** acquired ReadSoft; since then a lot of people have left (including Malchow and Göbel),” said Malchow. “Last year, Ralf and I got together with Dan and discussed big picture trends like capture platforms, robotics, and crowdsourcing. It soon became crystal clear that there was an opportunity to join forces. There was great chemistry among all the protagonists.

“We have one of the most experienced

teams in the capture market—especially for BPO and shared services operations (which is where Foxray’s software was targeted and where ScaleHub is initially marketing its services). And Dan and his team developed from scratch a cloud-based software platform to manage crowdsourcing.”



Olaf Malchow, CEO,
ScaleHub

As mentioned, ScaleHub is targeting the BPO and shared services markets. Shared services are like in-house BPO operations, in that they might handle document processing for several departments. ScaleHub is focused on shared services within the financial services and insurance markets. “In addition to shared service operations, which are typically run by large enterprises, there are approximately 10,000 BPOs worldwide that are interesting to us,” said Göbel.

Most of ScaleHub’s target customers are already running high volume capture operations. “We are offering them higher automation and better scalability,” said Göbel. “They can always scale up by adding more hardware and software, but when it comes to verification and data completion, you need humans and they are not always so easy to add. Crowdsourcing makes it much easier and typically less expensive to scale personnel. And adding crowd workers through ScaleHub does not require weeks or months of training.”

ScaleHub actually has two lines of business it is pursuing. “One is managed services, which leverages our

crowdsourcing operation and can also leverage automated recognition technology, depending on the customer requirements,” said Göbel. “Basically, the customer signs a contract and agrees to pay us a certain amount per field captured, and we deliver the results in a pre-defined timeline



**Ralf Göbel, COO,
ScaleHub**

with a pre-defined quality. We operate in compliance with security standards like the Safe Harbor in Europe, as well as U.S. regulations. Managed services is our strategic direction and that is where we expect to earn most of our money.

“Our other segment, which we think has synergy with managed services, is providing consulting and capture solutions. We have a very experienced consulting team and a good relationship with Lexmark. Lexmark delivers its

XBOUND software [foxray’s flagship product], and we provide implementation and training services. There is also the opportunity to introduce managed services into these environments.”

How’s it work?

ScaleHub’s managed services sit on a cloud, but there is “connector” that runs on its customers’ servers. “The connector communicates with our job portal,” said Malchow. “Our job portal connects to our job dispatcher which manages our integration with Amazon mTurk [a crowdsourcing portal run by Amazon].”

The process starts with the customer and/or ScaleHub defining the fields they want captured from forms. “Before the document images leave our connector, they are cut into snippets, which are delivered to our job portal on the cloud,” said Malchow. “They are delivered through a secure FTP transfer, and in best case scenario, we don’t see the whole document.”

“The portal enables us to manage our various jobs,” explained Göbel. “It enables us to see how far along they are to being completed, what percentage of the fields have been assigned to mTurk users, and other details. Through the job dispatcher, we automatically are able to publish jobs on mTurk. mTurk users can see job descriptions and decide to accept or skip them.”

ScaleHub uses crowdsourcing to complete a variety of capture tasks. These include data/field capture, OCR verification and completion, empty field detection, document classification, and table extraction. “Each job contains a clear description of what has to be done, as well as examples of how the data should be captured,” said Göbel. “We found that it’s important to have a good description if you want high quality work. We also provide training for specialized jobs like capturing German language documents, for example, where characters can differ from English.

“We’ve found that people like simple jobs, like checking if

Document Imaging Report

Business Trends On Converting Paper Processes To Electronic Format

DIR is the leading executive report on managing documents for e-business.

Areas we cover include:

1. Document Capture
2. Image Processing
3. Forms Processing/OCR/ICR
4. Enterprise Content Management
5. Records Management
6. Document Output
7. Storage

DIR brings you the inside story behind the deals and decisions that affect your business.

Vol. 26, No. 13



Editor: Ralph Gammon
4003 Wood Street
Erie, PA 16509
PH (814) 866-2247
FX (412) 291-1352
ralphg@documentimagingreport.com

DIR is published 23x per year, on the 1st & 3rd Fridays of the month, by:

RMG Enterprises, Inc.
4003 Wood Street
Erie, PA 16509
PH (814) 218-6017
<http://www.documentimagingreport.com>

Copyright © 2016 by RMG Enterprises, Inc. Federal copyright law prohibits unauthorized reproduction by any means including photocopying or facsimile distribution of this copyrighted newsletter. Such copyright infringement is subject to fines of up to \$25,000. Because subscriptions are our main source of income, newsletter publishers take copyright violations seriously. Some publishers have prosecuted and won enormous settlements for infringement. To encourage you to adhere to this law, we make multiple-copy subscriptions available at a substantially reduced price.

Subscriptions: \$597 (electronic) or \$670 (paper) per year.

there is a signature on a check. We pay Amazon and Amazon pays the mTurk users. The fees are based on several parameters—if it's machine or handprint, for example, or the complexity of the field—is the data just numbers or a full alphanumeric set?"

Dubiner said that mTurk is like any labor market. "At the end of the day, if you want to increase the speed at which you are getting things completed, you increase the fees you are paying, and our software kind of controls that automatically," he said. "We can also alert mTurk users with e-mails that there is a task that needs to be completed quickly. When you raise awareness through a proactive initiative like that, the [labor] market tends to respond.

"Also, we've learned that the usability of a task is important to keyers. If they are working fast and smoothly, they will like it better regardless of the pay. That said, we make sure not to abuse crowd workers. We never go below one cent per task, for example."

Exactly what a job entails varies depending on the requirements of the customer. "Some will have double blind key entry, and for other jobs we might leverage OCR results," said Göbel. "If there is a mismatch between two data entry efforts, the snippet will go to a third operator, who could be connected to a private cloud. A private cloud could be a more secure operation possibly within the BPO, where the keyer might be able to see a larger snippet that would provide context. Basically, we can adjust our workflows depending on the requirements of the customers.

"Once data is captured, it's sent back to the customers and integrated directly into the data models within their capture platforms. If there are fields that are illegible, they are sent back as empty fields and entered into the user's capture exception queue."

ScaleHub's goal is to not only increase the automation in its customers' current document capture operations but to also get them to increase the types and volume of documents they are capturing. "Most current capture operations are focused on machine print," said Malchow. "That's because BPOs and shared services centers don't have any way to capture handprint or written data. We want our customers to look at additional document types they might be able to run through their capture operations by adding our managed services component."

Off to good start; more to come

ScaleHub is being funded by its principals. "We

don't have any venture funding or credit/loans," said Malchow. "It is important for us to remain independent."

Initial sales appear to be brisk, as Malchow told *DIR* that the ScaleHub already had 15 customers.



**Dan Dubiner, CTO,
ScaleHub**

"Most of them are larger enterprises," he said. "The majority are XBOUND users, but we've also done integrations with Vinna [capture software from European ISV **Skilja**, see *DIR* 5/27/16]. Our vision is to help customers with any capture platform step into crowdsourcing solutions. It's very important for us to form partnerships to accomplish this." (ScaleHub's

other current partners include **Eucon**, a solutions provider which specializes in claims management, and **TCG**, a document capture and processing specialist.)

ScaleHub is prepared to process electronically generated documents, as well as images of scanned paper. "We are seeing volumes of electronic document formats like PDF, e-mails, and e-docs from ERP systems growing much faster than paper," said Malchow. "Every new mailroom project we enter is connected to e-mail. No matter what channel a document comes from, we can address it. For example, we can send snippets of PDFs to the crowd."

Malchow concluded that leveraging crowdsourcing is the next step in the digital transformation for capture users. "Current capture operations are like the assembly line that Henry Ford introduced in the 1930s," he said. "The traditional assembly line doesn't take into account the digital transformation being brought on by technology like the Internet and the cloud. Our vision is to help customers with a capture platform take the next step and involve the crowd in their production processes."

ScaleHub has deployed its software on Amazon Web Services (AWS), which helps it scale its offering worldwide. "We are the only vendor we are aware of offering these type of crowdsourcing services to customers worldwide," said Göbel. "Our approach is multi-lingual and working with AWS enables us to deal with localized compliance requirements."

"In addition," finished Malchow, "we are planning to continue developing more services beyond capture. There are some interesting things we think we can do with our crowdsourcing platform."

For more information: <http://www.scalehub.com>