PYRAMIDSOLUTIONS

VISIONARY SOLUTIONS ▲ EXCEPTIONAL RESULTS

Best Practices for Paper-Based Form Design

For Improved Document Capture ROI

Author: Scott Maloney, Senior Project Manager

Published September 2017

TABLE OF CONTENTS

In recent years, online activity exploded, but despite all of the rosy predictions about how we will become a paperless society, we still generate more paper than ever!

So, what is getting in the way of a faster movement to a paperless society?

A lot of interaction with customers is still driven by paper-based forms and there are challenges that come with this type of interaction.

The person filling out a form is out of your control. No matter how well you design the form there will always be responses that cannot be read automatically. You can encourage the form users to write neatly and keep their responses within the allotted space, but there will always be people who don't read instructions (or don't care), and assume the form will be read and processed by a human, not by a computer.

They do things like write a character by mistake then draw a big "X" or scribble over it to "delete" it. In a day and age where more and more communication happens online with a keyboard, mouse or touch screen, handwriting is getting more and more messy.

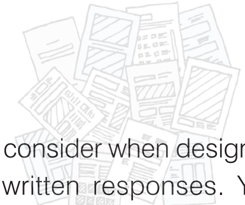
However, this challenge creates an opportunity for organizations to re-design forms so they can optimize their capture investment.

Not only will this make the process easier, but as an organization redesigns more and more forms, it will reduce the most expensive part of the process: manual intervention required during validation to make corrections to captured information before exporting the document to the corporate repository for processing.

This white paper seeks to provide meaningful guidance to companies that want to reduce the human capital costs associated with manual verification of incoming document images by re-engineering their forms.

By following these guidelines, you will notice more accurate written information, a smoother capture process, and a cleaner corporate repository.

PREPARING FOR FORM DESIGN



There are many factors to consider when designing a form to collect handwritten responses. Your target audience needs to easily understand the form which means where the user is supposed to write their responses is obvious.

The industry average for intelligent character recognition (ICR) is about 70 percent. You can never expect 100 percent accuracy, but shooting for 85 percent is considered good (even though it's still 15 bad characters out of 100). Good form design planning can usually exceed the 70 percent threshold.

When designing new forms, try to maintain a balance between aesthetics and readability. Every form is unique and should be approached with a cost-benefit analysis. The goal of form design is not to eliminate all data entry but to create cost-efficient forms. There will be at least a small degree of verification required on any form.

Take time to plan how to design a form. You should identify the field types required for layout prior to development. The specific purpose for each form plays a significant role in how to develop it.

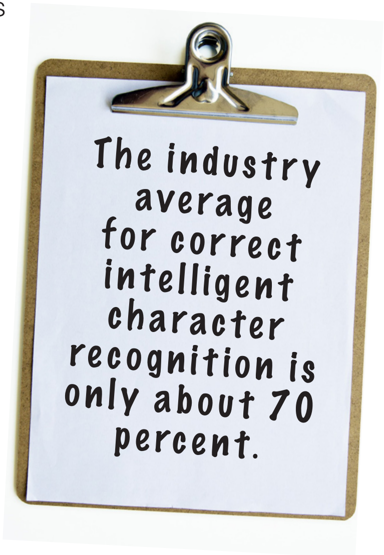
Readability is important when designing any new form. Always keep in mind the intended audience and the form's intended purpose. Simple user instructions can significantly improve recognition rates.

- Indicate on the form that a computer is going to process it
- Stress the importance of writing plainly, carefully and clearly

- Ask for block letters on the form
- Put instructions in bold at the top of the form
- Show examples such as how to write an “A” or “2”

Planning Steps

1. Identify and meet with the form owner. If this is for an existing form, determine whether you can redesign the form.
2. Obtain copies of all versions of an existing form.
3. Identify the goal or intended purpose of the form and the form’s audience.
4. Work with appropriate technical personnel (DBA, application developer, etc.) and representatives from the business (business analyst, subject matter expert, etc.) to obtain a list of all required fields and field properties.
    - Field properties should include maximum field lengths, field types, and required fields.



The industry average for correct intelligent character recognition is only about 70 percent.

5. Identify the capture software features and functionality that the form will go through.
6. Involve all parties that are directly connected to the form manipulation before it gets to the OCR system.
    - For example, a customer may send a filled out form to a specific business department where the personnel will then review, and possibly add additional data/stamps/other markings. If the customer fills out the form neatly, but the department personnel place stamps on important data, OCR accuracy will be affected dramatically, and will not be as good as it should be. The personnel above should be trained to know this, and perhaps the form design should include fields or areas named “Internal Use Only” where they can place any stamps, writing, etc., so it doesn’t interfere with other clean data on the image.

Questions to Consider

1. Who is the audience? Do they require special consideration like larger fonts or extra space between fields?
2. Who owns the existing form? Does this prevent redesign?
3. Are aesthetics an especially important feature for the form?
4. What is the form’s intended purpose or goal?
5. Approximately how many fields will be on the form?
6. Are there any areas on the form that require free form entry?
7. Do previous versions of the form exist, or is this a new form?
8. Can the layout of previous forms help with the template for the new form?
9. Do you have to capture data from an existing form until you can implement a new form?
10. What is the expected submission volume of the forms?
11. Do the forms require peak processing periods (e.g., certain months of the year will generate huge volume while other months will be comparatively light)?
12. Can you break a packet of forms into multiple form types?
13. Does each form have a unique identifier to link the documents together after the capture process?
14. Is the form a multi-page document? Do the pages have to be linked together after processing?

15. Will a double-sided form be used? Keep in mind the quality of paper can impact recognition accuracy in double-sided forms. Standard 20# copier paper works well for single sided forms and 24-28# paper works well for double-sided forms to prevent the back-side content from bleeding through when scanned. Fields may also be offset to ensure that any bleed-through content from one side will not interfere with the field recognition on the other.
16. Will you print the form internally or at an outside vendor?
17. Do you know the field requirements for every field on the form? Where can you obtain this information?
18. Which parties will enter data on the form? (i.e. customer, internal department ((received stamping, other, etc.))

FIELD TYPE CONSIDERATIONS

Proper layout of the printed response areas significantly impacts the accuracy of hand-printed content recognition. A common mistake in field design is to provide a free form area for response (i.e. a simple blank line where people should write). Without any character restraints, people will run characters together and write in cursive, on top of the line and in multiple lines which significantly impacts recognition accuracy.

A well-designed form has a defined response area for each character, encouraging character separation. Character boxes, comb lines and choice fields yield the highest degree of character recognition.

A form can use image zones to capture and recognize free-form text, however these zones require significant testing and will never achieve the caliber of recognition levels associated with constrained fields. Verifiers often have to correct or input handwritten information as free-form text into free-form zones on a form.

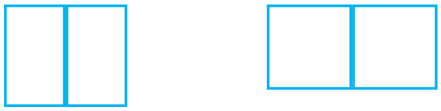
Character Boxes

Character boxes are the best method to encourage character separation. A good character box design allows users to write characters completely within each box. Unfortunately, many forms contain boxes that are too small and too close together. People often can't write small enough to keep an entire character within each box. Pencil lead creates strokes that are usually wider than pens, making it even harder to constrain the character.

The following are some general guidelines:

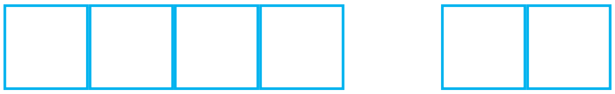
Each box should be square in shape, and the line width should be 1 pt. Thicker lines run the risk of leaving partial lines behind when the document goes through the image enhancement process. Rectangular boxes with a height taller than the width can make the user feel like they need to squeeze their characters into the space. This often results in characters written in a compressed vertical form, reducing accuracy. A square shape encourages wide, formal characters.

Narrow boxes Square boxes

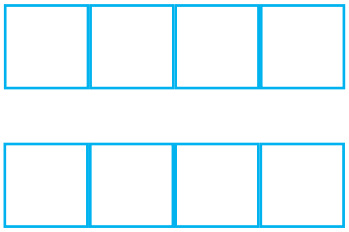


Create single character response locations, such as for Male ("M") or Female ("F"), in a single box separated from other responses.

Separate individual fields with enough space to easily identify where one field stops and the next one starts. I recommend **spacing of at least 1.5 box widths** to prevent users from interpreting the space as a valid character location.



Be sure to separate rows of fields stacked vertically by at least one-half the height of an individual box.



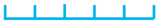
You can print boxes with either solid black lines or dropout colors, depending on the scanning and forms processing technology. Software-based form dropout removes the boxes from the image after scanning.

Some forms processing systems require dropout colors when printing forms. For example, a form might be in red ink and a red bulb in the scanner eliminates the red content during capture.

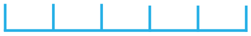
Comb Lines

Comb-style fields consist of short lines between each character. Although a commonly used format, people rarely align the characters within each space. The spacing between the vertical lines is often too close together, making it almost impossible for the average person to stay between the lines. If using comb lines, provide plenty of space between the vertical tick marks. Make the tick marks tall enough to encourage people to write between them. A vertical height of at least half the height of the expected character is usually sufficient.

Example of a poor comb line.



Example of a good comb line.



Prior to developing a form, obtain a copy of the database schema for the backend system. You will need to specify field lengths, data types and required fields. It is important that these attributes are consistent with the backend environment. Field lengths can be shorter than the limits defined in the backend system, but not longer.

Every field type has a confidence level (%). **A confidence level measures how certain the read of information on the field must be before bypassing verification.** You can set confidence levels high at the beginning and cautiously lower them as verifiers see over time that the information reads are accurate and seldom require manual correction.

Choice Fields

Choice fields consist of a list of options. Choice lists can be set up for a single or multi-choice response. The form recipient chooses from the options by selecting the appropriate response mark. Examples of response marks include bubbles, boxes, brackets or lines.

Gender
O Male
O Female

		A	B	C
1	O	O	O	
	2	O	O	O
	3	O	O	O
	4	O	O	O

Bubble fields always yield the highest degree of accuracy. The choices to select on the form may be simple, but keep in mind you can configure the field to export detailed values based on the option selected on the form.

The form designer can tie each response mark to a specific value. You can customize the range of values, such as A-Z; 1-5; A-E; 1-9. Choice fields are often used in multiple choice tests, surveys and standardized tests. They are extremely accurate, but may be inefficient for entering alphanumeric and alphabetic text.

Optical mark recognition (OMR), sometimes known as “mark sense,” is the analysis of form locations to determine if a mark is present. Whether you use oval bubbles, square boxes, open brackets or signature blocks, be sure the area is large enough for the user to easily mark within the designated area.

Common OMR field design errors include making the box or oval too small for people to easily mark within the zone, or printing the boxes or ovals too close together, resulting in more than one space containing the mark written by the user.

Bubbles between 10 and 14 points in height work the best. I recommend using a **capital “O” in an Arial font** (do not use the Times New Roman font). A non-oval or perfectly round circle is somewhat harder to fill completely, so respondents will subconsciously be compelled to fill in the circular bubbles more completely and neatly, leading to better recognition rates.

It is possible to place numbers or letters inside each shape, however the number/letter must be as small and light as possible. Dark, thick or bold characters may cause false positives because the OMR interprets the space as filled during scanning.

Spacing

Allow at least 3/8 inch between any text, lines or graphics on the form and any bubble areas, OCR text or barcodes. Minimize potential errors by separating bubbles from one another by at least two character spaces.

Try to **stay away from using lines or boxes around or between the bubbles on the form.** If you find it necessary, consider making the lines a light gray or red that will drop out (completely disappear) during the scanning process. Keep in mind that if people try to make copies of forms with shaded areas, those areas will appear darker on the copies.

Some users will circle an OMR response instead of filling in the box. Similar to character responses, providing clear instructions and example marks will significantly improve recognition results. Even great instructions will not prevent some people from marking a zone in error then drawing a big “X” in an attempt to make a correction. You will need to develop business rules to handle mark situations and manual key-from-image operations to determine user intent and make corrections.

Image Zone

Image zones are extremely flexible and useful when designing forms. Image zones are for capturing information that is best entered manually. Verification operators have to read the written response and type it into the capture software.

Email Address

Other Comments

Keep in mind that image zones are never as reliable as character boxes for OCR and ICR. Use image zones for any of the following purposes:

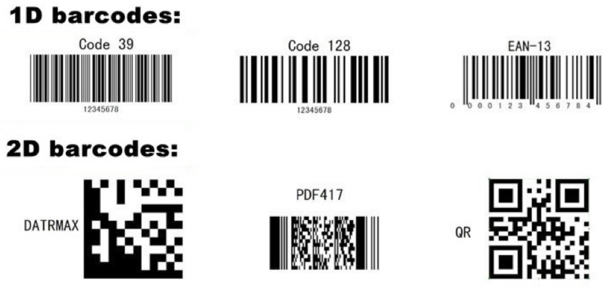
- Reading hand and machine print
- Reading or printing barcodes
- Calculating the percentage of an area filled
- Saving an area of a form as an image or BLOB

Within image zones you improve recognition by specifying font type and if the expected content is going to be machine printed versus hand printed. Image zones combine the reliability of manual data entry with the efficiency of automated recognition and data validation. For example, the verifier can draw a band around an area to display it for verification and to allow for manual entry into the appropriate field(s).

A large image zone can be set up for multiple handwritten lines of free-form text or when a set of fields is either difficult or inefficient to recognize. Input options are tied to the image zone. For example, a verifier may manually input a social security number that was handwritten into the middle of an image zone, or select a value from a choice list on the validation screen.

Barcodes

Barcodes help classify document or form types that are scanned. Barcodes come in 1D and 2D formats and can be horizontal or vertical on a form.



Barcodes should be at least 26 points in height. Similar to OMR bubbles, **leave at least 3/8 inch of space around your barcode.** Remember that barcodes may vary in length depending on what you are capturing. Leave room for the largest expected barcode both on the form and when the region definition is configured in the capture software.

FOCUSED RECOGNITION and DATA VALIDATION

Some fields are designed to allow only certain characters. For example, a date field may allow only digits, or only digits, dashes and slashes. A "Male/Female" field may only allow the characters M and F. **Ensure that your form contains instructions or examples for each field to ensure the user knows what characters are allowed.** You can factor the definition of allowable characters into the verification software so it flags invalid characters for review.

Low confidence data typically requires a "key from image" process to validate. This requires displaying suspect characters or fields to a human for manual data entry. Human interaction is the most expensive part of any data capture process, so any efforts you can take such as strong form design or additional image enhancement processes, will easily pay for themselves when compared to the cost of manual data entry.

OTHER OPTIMIZATION RECOMMENDATIONS

Limit the List of Acceptable Characters for a Given Field

Expected characters can be alpha, numeric or alphanumeric. Specify whether the received text will include hand or machine print. Define which special characters you anticipate receiving on the form. For example, a telephone number will never include alphanumeric characters and should be limited to numeric. You may also want to identify the dashes within the list of expected characters if you did not use a constrained-print approach for the field.

Put a Form Identifier in the Same Location on Every Page

When scanning, indexing and verifying multi-page forms, you can improve the accuracy tremendously by putting a form ID in the same location on every page with sufficient white space around it. Format the form ID in a way that facilitates easy recognition of the pages. For example, form ABC-1 for page one, ABC-2 for page two and ABC-3 for page 3.

ABOUT THE AUTHOR

Scott Maloney has held various management positions in marketing, operations and information technology with financial services companies for more than 20 years. During his tenure with Pyramid Solutions as a Senior Project Manager, Scott has managed enterprise content management projects that resulted in the deployment of new solutions across a spectrum of state government, insurance and banking clients.

ABOUT PYRAMID SOLUTIONS, INC.

Pyramid Solutions develops products and innovative solutions for organizations in a wide range of industries – from financial institutions and insurance providers to automotive suppliers and industrial automation companies. We serve primarily as an Intelligent Automation company that specializes in RPA, Business Process Management, Content Management, Capture/OCR and Industrial Automation solutions including MES software and embedded software development. With more than 30 years of experience, we know a thing or two about automation.

To learn more, visit pyramidsolutions.com

REFERENCES

Scanlan, Rick. "Best Practices: Improving ICR Accuracy with Better Form Design." Published August 25, 2015. https://www.accusoft.com/whitepapers/best-practices-improving-icr-accuracy-with-better-form-design/.

Remark. "Form Design Best Practices." Accessed August 8, 2017. http://remarksoftware.com/support/office/form-design/

Verity. "Scannable Form Design Best Practices." Published April 2005. http://www.consiliumdcs.com/tools/spaw/files5/Forms%20Design%20Best%20Practices.pdf

PYRAMID SOLUTIONS

VISIONARY SOLUTIONS ▲ EXCEPTIONAL RESULTS