The extraction spec

Define exactly what to pull from a document, what to do when something's missing, and how you'll check the result.

Time **about 20 minutes** Updated **12 Nov 2026** Fill in online at **Learn › Worksheets**

HOW TO USE IT

1. Pick one document type and three real examples you're allowed to use.
2. List every field, its format and a rule for missing values.
3. Run the spec on all three examples and record what your checks found.

1. The document

Document type

How I'll get the text out, and what I checked in the conversion

2. Fields

What to extract

	Field	Format	If missing
Field 1			
Field 2			
Field 3			
Field 4			
Field 5			

Rules

3. Checks

Arithmetic or script checks

What I'll spot-check by hand

Test results

	What the checks found
Example 1	
Example 2	
Example 3	

Example answers

> The example is the fictional invoice from the lesson. Example 1's result is from our real run on 12 November 2026; the others are for you to fill in.

1. The document

Document type

> Supplier invoices (Word and PDF)

How I'll get the text out, and what I checked in the conversion

> MarkItDown. Checked: table rows intact; header row shifted down one, which is fine.

2. Fields

What to extract

	Field	Format	If missing
Field 1	supplier	text	null
Field 2	invoice_date	as written	null
Field 3	due_date	as written	null; never calculate from terms
Field 4	lines	item, qty, unit price, line total	empty list
Field 5	total	number, no £ sign	null

Rules

> Use only what the document says. If a field isn't stated, use null. Never calculate or guess a value. Reply with JSON only.

3. Checks

Arithmetic or script checks

> qty x price = line total; lines add up to subtotal; subtotal + VAT = total.

What I'll spot-check by hand

> Supplier name and invoice number against the original.

Test results

	What the checks found
Example 1	All checks passed.
Example 2	(your result)
Example 3	(your result)

A schema fixes the shape, not the facts. Check numbers with arithmetic and text against the original. Don't upload documents to a tool your organisation hasn't approved. Free to use and adapt for your own work.