Tickboxes

Yep, tickboxes. One of the issues for many real-life entities is data entry from paper forms (as opposed to the online forms mainly Internet-based entities: yes, yes, the physical world does still exist, and yes it is easier sometimes to get people to put crosses on paper than to lead them to the right website/ phone app). Specifically, they use paper because it's their best shot at getting lots of people to fill in data fields, but they then have lots of paper data that they need to digitise.

Options for this, rated from high to low operator involvement, low to high flexibility and low to high complexity, are:

- Use the forms as raw data, i.e. do all the analysis needed by counting up ticks in boxes etc by hand. This works, but takes time, and isn't easy to cross-check or add to the analysis used; it can however be optimal for small amounts of data where the analysis is needed quickly.
- Get someone to type in all the data from the forms. This takes time, but is usually a faster way to get relatively small amounts of data ready for analysis than spending time working out how to do the digitisation.
- Scan the forms whole into the system. This captures the data online, but is no better for analysis than the first option.
- Scan the forms then use OCR and image processing to capture the data on each form. This works for some limited types of data: typed data, neatly filled-in tick-boxes, carefully-spaced capital letters (a la postcode reader), but doesn't capture free text and may have problems with messy inputs.
- Scan the forms and use OCR/ image processing to capture as much data from the form as possible, then use the operator to cross-check the captured data (e.g. by online comparison between the original form and the results) and input any free text or other difficult-to-read data.

I'm looking at the last option for some IT4Communities stuff. As with most autonomy, it makes sense to use the human's and machine's strengths together, to work comfortably somewhere away from both excessive operator workload (low autonomy) and massively complex systems (high autonomy).

I could write a segment-then-find-boxes-then-check-their-occupancy image processing subsystem, but I suspect that because this is a relatively common problem, someone has probably already done this. I'll attempt to write the system out of sheer curiosity if I can't find one, but meanwhile places I'm starting to look include:

- census data processing
- exam paper processing
- Search for freeware paper form processing

Helpful references include:

- Census - choosing the right paper-based data capture method.
- Omnipage
- FormReturn recognition guide