



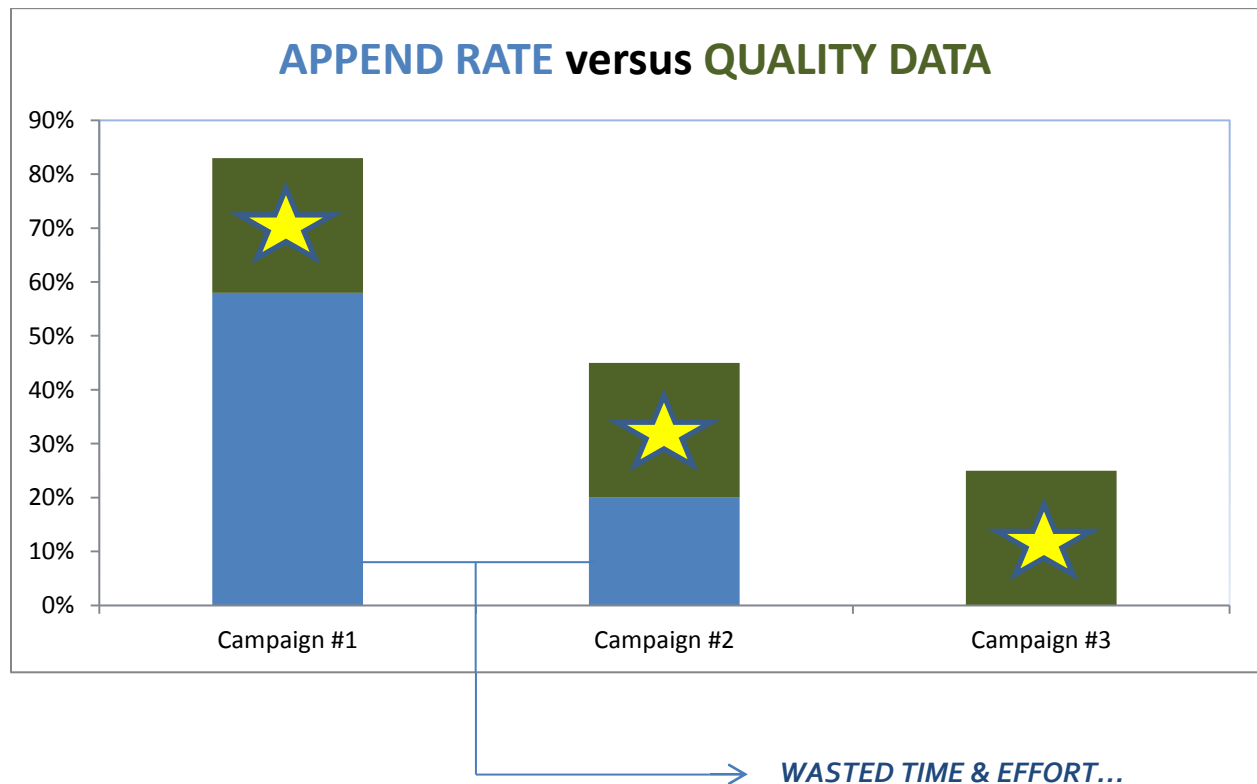
**INFORMATION EVOLUTION**  
Reliable, Experienced, Affordable Human Bandwidth.

# A CASE STUDY IN MANAGED CROWD SOURCING

## HIGH-QUALITY DATA IS ALL ABOUT SAVING *TIME* AND *MONEY*

### PURPOSE

The purpose of this document is to compare and contrast process methodology between three different campaigns to append URLs and email addresses to an Italian company database.





## INFORMATION EVOLUTION

Reliable, Experienced, Affordable Human Bandwidth.

### DEFINITIONS

**Campaign #1:** Run through Amazon Mechanical Turk (first initial run)

**Campaign #2:** Run through Amazon Mechanical Turk (general task instructions were rewritten and clarified using lessons learned from Campaign #1)

**Campaign #3:** Run through Crowd Control's software

- IEI used CC's URL validation entry which only accepts properly formatted URLs
- IEI used CC's email validation that accepts only properly formatted email addresses [xxx@xxx.com]
- IEI submitted to CC a list of text strands for the URL field to automatically reject which included applegate, europages, facebook, onsource, youtube, hoovers, dnb, paginegialle, yell.co.uk
- IEI had CC set up URL pings to confirm that a website actually exists and is functioning to eliminate dummy URLs submitted by workers
- Task instructions were also rewritten using lessons learned from both previous campaigns

### RESULTS

Metric	Campaign #1	Campaign #2	Campaign #3
URL append rate	83%	45%	25%
Email append rate	64%	50%	25%
Avg. time per assignment	1:13	1:33	1:33
Effective hourly rate per worker	\$0.98	\$0.77	\$1.55
Initial submission quality	Very poor	Slightly better	100% accurate after IEI QC check
IEI QC review time	Dozens of hours	Several workdays	A few hours
Final deliverable quality/accuracy	LOW	MEDIUM	HIGH



## INFORMATION EVOLUTION

Reliable, Experienced, Affordable Human Bandwidth.

### ANALYSIS

- **Campaign #2 produced a 20% lower URL return rates due to false positives in Campaign #1**

Campaign #2 consisted of a more rigorous removal of *false positive* URL entries. For example, a quick search of Campaign #1 data shows following false positive web sites: YouTube, Facebook, OneSource, Coop Biz, EuroPages, and several others. Other false positives occurred because of the following two situations: 1) An alphabetical review of URLs shows that companies with similar names but that were different companies had the same URL listed; 2) Some URLs that were accepted were simply the full company name placed between “www” and “com”. Also, in several instances, Campaign #1 allowed parent company web addresses as acceptable.

Campaign #2 vetted out any and all false positive web sites via find and replace methods for the above commonly used false positive URLs as well as several others discovered during the data vetting process. Campaign #2 also used a “triple-check” formula in Excel to highlight inaccuracies from one data set to the next which caused more and more URLs to be thrown out as false. Campaign #2 was also reviewed in a more rigorous fashion to eliminate duplicate URLs – e.g., when the same URL was used for two or more different companies. Campaign #2 also threw out parent company URLs and listed NIF (no information found) when the subsidiary did not have its own independent web page.

The methods of Campaign #2 netted fewer URL appends due to more rigorous data vetting using lessons learned from Campaign #1.

- **Campaign #2 produced only a slightly higher percentage of appended emails addresses (a 5% increase) than Campaign #1**

Campaign #1 allowed for the collection of a secondary email addresses which resulted in inconsistent data gathering for the primary email address. Campaign #2 also had a number of instances in which a subsidiary company with no independent website did in fact have a direct email contact site via a parent company website; as a result, there was a distinguishing factor between Campaign #1 and Campaign #2. As previously mentioned, Campaign #2 did not accept parent company URLs, however, subsidiary email information found within parent company websites were included. As a result, there may have been instances in Campaign #1 where if a company URL was returned as “NIF” then the worker would automatically put “NIF” in the company email.

NOTE: It should be noted for company email addresses, the address is often simply info@companyname.com. Because so many companies actually have “info@” email contact addresses, it is difficult to assess which “info@” email addresses are valid and which ones are dummied up by workers. The industry should investigate opportunities to validate email addresses in much the same manner as we are validating URL addresses. However, issues of spamming must be considered as part of the plan to validate email addresses and may actually preclude this endeavor because of the danger of being blacklisted.



## INFORMATION EVOLUTION

Reliable, Experienced, Affordable Human Bandwidth.

- Campaign #3 produced lower URL and email append rates and cost more...but data quality was *excellent...*

Campaign #3 produced a fully automated URL and append rate of about exactly 25% with no need for manual confirmation or rigorous QC by IEI. A fast, initial QC of about ten records produced a 100% accuracy rate.

The hourly rate increased while the append percentage declined because of the near total practical absence of false positives. While the append rate “decreased,” what actually happened was that the actual accuracy of the appended data increased dramatically. IEI also feels that overall, workers feel more compelled to enter some data, even if it is incorrect, than to list “N/A” for fear they will not get paid, when in fact the opposite is true. IEI feels that the project setup made it very clear to turkers that “N/A” is in fact valid and that we would rather have less, more accurate than more, less accurate data. This is exemplified by the often-used analogy of a “shotgun” versus a “laser-sited rifle” approach. It can also be likened to the difference between finding what you are looking for at a garage sale versus using eBay.

## POOR DATA COSTS CLIENT & VENDOR *TIME AND MONEY*

Poor data pollutes the task at hand by expending unnecessary time and money. Every time you go a URL that doesn't exist, send an email that gets returned, or make a phone call to an incorrect number, you are wasting time and money. Productivity plummets and confidence erodes with every minute spent chasing poor data.

## HIGH-QUALITY DATA SAVES *TIME AND MONEY*

High-quality, accurate data is the key to any successful data-led endeavor. Every time you go a URL that DOES exist, you are making that connection. Every time you send an email or make a phone call that goes straight to the intended target, you are making that connection. Not only are you making that connection, but you are making it in the most efficient and cost-effective manner possible.



## INFORMATION EVOLUTION

Reliable, Experienced, Affordable Human Bandwidth.

### APPENDIX

#### Using Language Qualifiers

IEI attempted to use an Italian language qualification test but received a response rate so low as to negate that specific qualification.

Amazon Mechanical Turk does not currently collect language proficiency information from turkers. The closest one can come to getting language specific is by noting the country a worker currently resides. This gives no guarantee of their specific language proficiencies as an American could easily reside in Italy, an Indian could easily reside in England, etc.

Although it did not pan out this time, the Italian language qualifier test was an excellent experiment and IEI has an excellent working template if this can in fact be utilized in the future.