Non-survey sources of data to supplement the traditional measurement of e-commerce and the digital economy

Working Group on Measuring e-commerce and the Digital Economy, Second Meeting
ABOUT CETIC.BR

15 YEARS MEASURING DIGITAL TRANSFORMATION IN BRAZIL

Source: NIC.br / Cetic.br (2020)
The Regional Center for Studies on the Development of the Information Society (Cetic.br) has the mission of producing ICT statistics.

1. KNOWLEDGE CENTER ON ICT-DATA PRODUCTION
   - ICT public statistics
   - Statistics for the SDGs
   - Disaggregated data
   - Qualitative studies

2. CAPACITY-BUILDING METHODOLOGIES FOR THE PRODUCTION AND USE OF ICT-STATISTICS

3. LABORATORY OF IDEAS & METHODOLOGICAL INOVATION

4. MEASUREMENT OF IMPACTS OF ICT IN SOCIETY, PRODUCTION OF STRATEGIC RECOMMENDATIONS & POLICY BRIEFS
REGULAR PRODUCTION OF ICT INDICATORS

MONITORING POLICIES AND INTERNATIONAL AGENDA

METHODOLOGICAL FRAMEWORKS, POLICY BRIEFS & COUNTRY REPORTS

ICT survey projects

Sectoral studies

MEDICIÓN DE LA SALUD DIGITAL

Practical guide for implementing surveys on ICT use in primary and secondary schools

UNESCO’S UNIVERSAL ACCESS INDICATORS
ICT DATA PRODUCTION ECOSYSTEM & MEASUREMENT PROJECTS AT CETIC.BR

DATA USERS
- Governments
- International Organizations
- Civil Society/NGOs
- Private Sector
- Academia/Research
- Media
- General Public

DATA PROVIDERS
- Administrative Data
- Survey data & census data
- Official agencies & institutes
- Official statistical institutes & research centers
- Ministries

PRIVATE DATA PROVIDERS
- Organic Data (Big Data)
- Mobile
- Crowdsourcing
- Online traces
- Social media
- Financial transactions

DATA SOURCES
- Survey Data
- Administrative Data
- New Data Sources (Big Data)
- Qualitative Data

FRAMEWORKS FOR DATA SHARING
- Technical
- Legal
- Institutional

TRUSTED DATA SHARING PLATFORMS

METHODOLOGICAL APPROACH
1. RESEARCH PLANNING
   - Project Design
2. DATA COLLECTION
3. DATA PROCESSING & MANAGEMENT
4. DATA ANALYSIS
   - Field Operations and Quality Management
   - Framing & Registries
   - Microdata
5. DATA USERS
   - Publications
   - Data Visualization
   - Microdata Portal
   - Capacity Building
   - Policymaking & Research
7. DATA DISSEMINATION
8. PROCESS MANAGEMENT & EVALUATION

PARTNERSHIPS & COOPERATION
Web scrapping for measuring e-commerce
PLAN

- Present Cetic.br/NIC.br’s experiences in the use of alternative methods of data collection to measure the digital economy. Which includes:
  - Integration of traditional survey databases and data collection on the web;
  - Use of private web scraping databases in conjunction with administrative records;
  - Use of traditional survey databases in conjunction with administrative records and data collection on the web.
ICT ENTERPRISES SURVEY

SURVEY OBJECTIVES
The primary objective of the ICT Enterprises survey is to measure the ownership and use of information and communication technologies (ICT) in Brazilian enterprises with 10 or more employed persons.

TARGET POPULATION
The universe for the ICT Enterprises survey consisted of all active Brazilian enterprises with 10 or more employed persons.

COVERAGE
• National

SAMPLE SIZE
• 7,000 enterprises

DATA COLLECTION
• Computer Assisted Telephone Interviewing – CATI
Brazil is the sixth country among G20 economies in country code Top-Level Domain registres. In April 21st there was 4,64 millions registers under .br.
Experiment 1 - Measuring e-commerce using Web Scraping

BUILDING A PREDICTION MODEL
Web scraping process
Survey Data
Dictionary

- website1.com.br
- website3.com.br
- website4.com.br
- ...
- websitesn.com.br

PHASE 1

PRODUCING E-COMMERCE INDICATORS
Sampling (Domain Name System Frame) (DNS)
Web data collection
Statistical Model

PHASE 2

Dictionary
DNS
Domain Name Database
Web Scraping Process
Data Analysis & Report
FINDINGS AND CHALLENGES

- The models developed did not provide accurate estimates for the selected indicators;

- The difference may be related to the sites not collected in the survey;

- The variation in the number of websites may be due to the extinction of enterprises, M&A, and other reasons that can contribute to the changing or termination of the websites;

- Many enterprises did not inform their selling websites, but the institutional ones;

- The search for a better model continues. Until we can estimate properly the indicators, we can not move to phase 2 of the project.
Experiment 2 - Combining web scraped data with administrative records to predict e-commerce indicators

- Use of different big data sources;
- Identify the domains registered (Registry of domains) and link to the private database of scrapped web sites.

<table>
<thead>
<tr>
<th>Source</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dataplayer</td>
<td>2,263,397</td>
</tr>
<tr>
<td>Registro.br</td>
<td>804,294</td>
</tr>
<tr>
<td>Cetic.br database</td>
<td>818,277</td>
</tr>
<tr>
<td>Cetic.br identified</td>
<td>530,844</td>
</tr>
<tr>
<td>.br Domain dataset</td>
<td></td>
</tr>
<tr>
<td>.br websites scrapped</td>
<td></td>
</tr>
<tr>
<td>.br websites scrapped and classified (online presence)</td>
<td></td>
</tr>
</tbody>
</table>
Experiment 2 - Combining web scraped data with administrative records to predict e-commerce indicators

- Multiplicity of websites per local unit;
- Multiplicity of website per company (e.g. one single company linked to 1,884 websites).

Cetic.br identified database 530,844
377,291 identifiers of local unit
374,831 identifiers of companies
FINDINGS AND CHALLENGES

National Registry of Legal Entities linked to multiple websites:

- Some companies are active in administration, construction or simply have domain registrations that are used by other companies;

- The website must be associated with the company actually "using" the website as a form of online presence or e-commerce, and not by the company that owns the domain.

Enterprises versus local units:

- It is common for a larger company to operate in several sectors;

- Some companies have multiple websites associated with its various business activities. This implies that the unique registry of the companies are linked to different websites, in reference to different activities.
Next Steps

Lessons learned and new developments
NEXT STEPS

Improve the accuracy of the prediction models
- Reduce burden to respondents;
- Development of real time indicators;
- Allow for new understanding of the digital economy.

Link as many databases as possible
- Integrate databases that are not suitable to our preliminary goals;
- Develop data collection using social media.

Face new challenges generated by the pandemic
- Improve the gain of knowledge on e-commerce;
- As digital maturity increases, big data will be more useful;
- Privacy legislation on web scrapping.
Experiment 3 - Use of traditional survey databases in conjunction with administrative records and data collection on the web

Companies with:
• Domains
• Web sites
• Social network pages
• Not found in the process

Web Scraping Process
Classification Process
Data Analysis & Report
Thank you!

www.cetic.br