IFC-ECB-Bank of Spain Conference: "External statistics in a fragmented and uncertain world"

12-13 February 2024

Web scraping as a detection tool in identifying Indonesian cross-border digital trade actors[1]

Aditya Wisnugraha Sugiyarto, Marlina Novita Uligoma, Hasudungan Paulanka Siburian, Dwi Cahyo Ardianto, Novi Ajeng Salehah and Detasya Avri Magfira,
Bank Indonesia

[1] This contribution was prepared for the conference. The views expressed are those of the authors and do not necessarily reflect the views of the European Central Bank, the Bank of Spain, the BIS, the IFC or the other central banks and institutions represented at the events.

Web scraping as a detection tool in identifying Indonesian cross-border digital trade actors

Aditya Wisnugraha Sugiyarto[1], Marlina Novita Uligoma[2], Hasudungan Paulanka Siburian[3], Dwi Cahyo Ardianto[4], Novi Ajeng Salehah[5], Detasya Avri Magfira[6]

Abstract

Massive technological advances have enabled the shift of the trade conducts, enabling the rise of digital trade that gradually replacing the physical exchange of products. Accordingly, tracking the digital trade activities through a robust statistic is becoming more important, including the cross-border trade. A good cross-border digital trade statistics will help in understanding how such trade evolves and how it impacts the economy. However, there are challenges in compiling the statistics, especially in collecting the data. To identify whether a cross-border trade can be classified as digital, the products exchanged should be either ordered or delivered digitally. One approach for such identification is to identify whether the actors involved are digital trade players or not. If the actors are digital trade players, then transactions involving their products could be assumed to be digital trades. This study proposes such identification by leveraging web scraping techniques to identify and compile the cross-border digital trade actors in Indonesia. Web scraping provides the ability to efficiently collect digital trade actors from various online sources, including e-commerce websites. In this paper, we will explore how the techniques can be used to overcome data limitations in digital trade analysis, enlarging the breadth of data coverage analysed. We also discuss aspects of legality, scalability, data security and data quality of such techniques, as well as the importance of complying with international statistical standards. By using the proposed method, we find that advanced analytics (including web scraping) in cross-border digital trade statistics significantly boosts exporter scope, improves efficiency, and ensures data quality by minimizing human errors.

Keywords: digital trade, web scraping, data engineering, statistics

JEL classification: C82, E58, F14, O19

[1] Department of Statistics, Bank Indonesia, email: aditya_wisnugraha@bi.go.id

[2] Department of Statistics, Bank Indonesia, email: marlina_nu@bi.go.id

[3] Department of Statistics, Bank Indonesia, email: hasudungan_ps@bi.go.id

[4] Department of Statistics, Bank Indonesia, email: dwi_ca@bi.go.id

[5] Department of Statistics, Bank Indonesia, email: novi_ajeng@bi.go.id

[6] Department of Statistics, Bank Indonesia, email: detasya_avri@bi.go.id

Contents

1. Background

In the era of globalization and economic digitalization, developments in information technology have brought major changes in the way trade is conducted. Technological advances are facilitating changes in trading behaviour, with digital commerce gradually replacing physical exchange of products (Ezel & Koester, 2023). Indonesia, as a country with rapid economic growth, is also experiencing the same significant transformation in its economic structure, especially with more transactions are carried out online and more products are delivered digitally. Digital Commerce has become a critical component of the modern global economy, enabling broader access to international markets, speeding up business processes, and giving consumers more choice (Kraus et al., 2021).

The rapid growth of the digital economy also brings a number of challenges in recording the Balance of Payments. As an important tool in measuring a country's economic transactions with the outside world, the recording of the Balance of Payments must be able to reflect changes in transaction patterns due to developments in the digital economy. This concern is reflected in the growing literature of digitalization inside the IMF's Balance of Payments Manual (BPM). The currently developing BPM7 has a dedicated section that discuss the framework of cross border digital trade, which would address the importance of the impact of such issue.

Some of the specific challenges faced in recording the Balance of Payments in the digital economy era and particularly in digital trade are: i) measuring digital transactions, ii) definition and classification of transactions, iii) uncertainty in transaction values, iv) identification of the parties involved (actors), and iv) adjustment to rapid changes (OECD, WTO & IMF, 2023). To address those challenges, a good methodology is needed in the process of recording digital trade, especially those involving cross-border transactions.

Indonesian digital economy potential is also huge. Statistically, Indonesia's digital economy is the largest among member countries of the Association of Southeast Asian Nations (ASEAN). In 2021, the size of the digital economy in Indonesia is around 42% of the ASEAN digital economy. Indonesia is also one of the most attractive digital investment destinations. Total investment inflows into the digital sector were $4.5 billion in 2020 and $9.1 billion in 2021 (Google, Temasek, dan Bain & Company, 2022).

The strength of Indonesia's digital economy comes from the high level of internet penetration and the large population of the younger generation (dominated by the younger generations Y and Z), which will become more dominant in the coming years (Alisjahbana et al, 2020). Both the government and the private sector have launched various initiatives to support the growth of Indonesia's digital economy. These initiatives mainly come in the form of physical and digital telecommunications infrastructure, investments in start-ups (including micro, small and medium enterprises) and general initiatives to improve the ease of doing business (East Ventures, 2022).

Indonesia's digital economy showed extraordinary growth of 414% in 2017 to 2021. In addition, it is estimated that it will grow around 62% in 2021 to 2025. In just less than 10 years, Indonesia's digital economy will grow eightfold. This is different

from the economic slowdown (shown by GDP) that occurred during the Covid-19 pandemic and the possibility of gloomy prospects for the global economy in the coming years (IMF, 2023; Sapulette & Santoso, 2021).

Sectors related to mobility and technology show higher growth than other sectors. Sectors related to technology support sectors related to mobility, such as transportation and tourism. Additionally, digital technology has enabled less popular tourism destinations to reach new consumers. Digital technology also allows small businesses to reach new consumers through online platforms, even if resources are limited. The digital economy has played an important role in supporting Indonesia's economic growth and will continue to do so (Sapulette & Muchtar, 2023).

According to the Central Agency of Statistics (BPS) in 2023, Indonesia has reached a population of around 275 million people in 2022, with 52% of the population dominated by generations Y and Z. This growth in the number of young people is in line with the rapid adoption of digital technology in Indonesia. The digital payment sector has also experienced significant progress, including the emergence of Electronic Money (e-money), card payments, QRIS, and mobile banking. Apart from that, e-commerce and digital financial services are also key sectors in Indonesia's digital economic ecosystem.

In the context of cross-border trade, identification of digital trade actors becomes increasingly complex. Available information is often limited to company names, and with the high volume of transactions, it is difficult to ascertain whether a company is truly engaging in cross-border digital trade (OECD & IMF, 2017). This challenge is further exacerbated by the lack of structured data and the variety of information spread across various sources.

The importance of data in supporting policy and decision making has encouraged the development of big data analysis (Suominen & Hajikhani, 2021). The use of big data analytics provides the ability to analyze and interpret large and complex amounts of data. However, in the context of identifying digital traffickers, it is necessary to look for more innovative and specific solutions.

Web scraping, as part of advanced analytics, is emerging as a potential solution to overcome the challenges of identifying digital traffickers. This technique allows collecting data from various sources automatically and efficiently. As technology develops, web scraping can become a very effective tool for accessing information about digital traffickers, including their activities and characteristics (Rundel & Dogucu, 2020).

Web scraping has experienced rapid development along with the need to efficiently gather information from an increasingly complex web. This technique involves the use of algorithms and software to extract data automatically, allowing for more in-depth analysis (Rundel & Dogucu, 2020).

Web scraping can be used to gather information from various digital trading platforms, company directories, and other related sources (Rundel & Dogucu, 2020). Its potential in accessing and collecting relevant data can provide a better understanding of digital trading actors.

By understanding the growth of the digital economy, the challenges of identifying digital traffickers, and the role of big data analytics, the use of web scraping becomes increasingly relevant as an innovative solution. This is what

underlies the author to apply this technique to provide a more accurate and in-depth understanding of digital trade actors, helping to develop policies that are adaptive to the dynamics of cross-border trade in the era of the digital economy in Indonesia.

Thus, this article contributes in terms of providing additional digital trade literature amid limited literature related to cross-border digital trade assessments particularly in Indonesia. In addition to that, the authors provide overview and framework of how to use advanced analytics to improve the quality of digital trade statistics compilations.

2. Literature Review

2.1. Digital Trade Concept

Digitalization affects international trade on multiple levels, by changing the way goods and services are traded and by creating completely new, internationally traded digital products. Equally important, digitalization also has a significant transformative impact on many existing industries: by "shrinking the distance" between consumers and producers, and between producers, digitalization provides previously unimaginable access to new markets, especially for micro, small, and intermediate business entities (MSMEs).

One of the main concerns driving the demand for better evidence on digital trade is the perception that large parts of the economy and international trade are going unaccounted for due to digitalization (Ahmad and Schreyer, 2016; Corrado et al., 2021). Even if it is generally accepted that the current statistical framework is still appropriate for measuring international trade, the fact that digital trade is not visible in existing statistics hinders the ability to assess the impact of trade policies and may lead to the misperception that digitalization in the trade economy is not being accurately measured.

The statistical definition of digital trade is based on the nature of the transaction, not on the characteristics of the products traded or on the characteristics of the actors involved in the transaction. Based on an international standard statistical manual, namely the Handbook of Measurement Digital Trade, released by the OECD, WTO and IMF (2023), digital trade is defined as: "All international trade that is ordered digitally and/or submitted digitally". Based on the guidebook, the explanation of the digital trade concept is explained in the following conceptual framework (Figure 1):

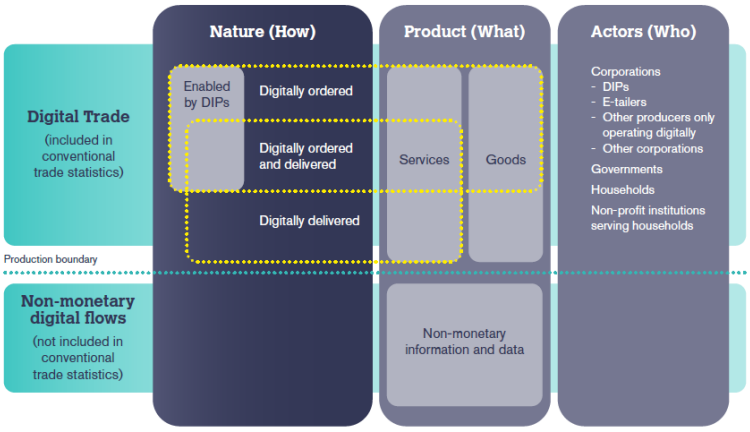


Figure 1. Conceptual Framework of Digital Trade (source: OECD, WTO & IMF, 2023)

Based on (see Figure 1), digital trade is divided into two types of transactions (how), namely:

a) Transactions for goods/services ordered digitally (digitally ordered)

Cross-border digital transactions ordered digitally are the international sale or purchase of a good or service, conducted over computer networks by methods specifically designed for the purpose of receiving or placing orders. The following supporting clarifications are provided to help identify digitally ordered transactions in international trade (OECD, WTO & IMF, 2023):

- For digitally ordered transactions, the payment and ultimate delivery of the goods or services do not have to also be conducted online.
- Digitally ordered transactions can involve participants from all institutional sectors.
- Digitally ordered transactions cover orders made over the web, 6 extranet or via electronic data interchange.
- Digitally ordered trade includes purchases of applications (apps) and in-app online purchases.
- Digitally ordered trade includes transactions via online bidding platforms
- Orders made by phone, fax or manually typed email are excluded from digitally ordered trade
- Offline transactions formalized using digital signatures are excluded from digitally ordered trade
- Each trade transaction should be treated separately. When a transaction is established via offline ordering processes, but subsequent transactions (or follow up orders) are made via digital ordering systems, the follow-up orders should be considered as e-commerce
- Trade transactions do not necessarily coincide with contracts. For a contract spanning several statistical periods and potentially involving multiple transactions, each transaction should be classified as

digitally ordered or not digitally ordered, reflecting the mode(s) of ordering initiated in the current period

b) Transactions for goods/services sent digitally (digitally delivered)

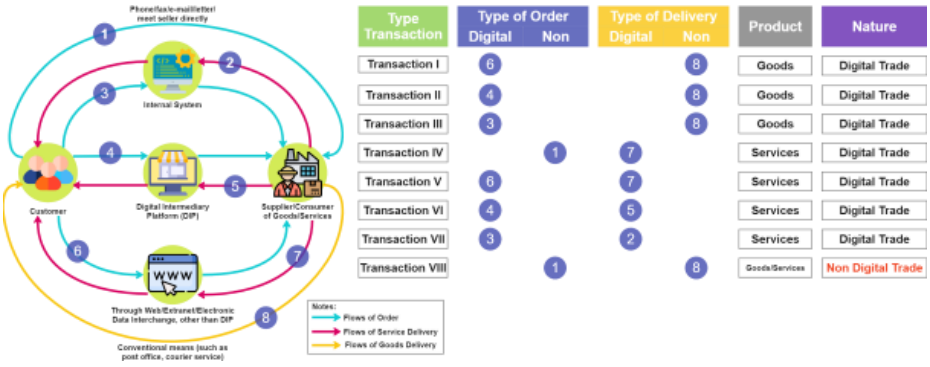
The second criterion for identifying digital trade is transactions which are "digitally delivered" and only cover services. Referring to the handbook, digitally delivered trade is defined as all international trade transactions that are delivered remotely over computer networks. The following supporting clarifications are provided to identify digitally delivered transactions in international trade (OECD, WTO & IMF, 2023):

- Only services can be digitally delivered
- Digitally delivered transactions can involve participants from all institutional sectors
- For digitally delivered transactions, the payment for and ordering of the services do not have to be conducted online
- Services delivered by phone, fax, video call or email are included in digitally delivered trade
- Digitally delivered trade includes services provided through apps
- Each trade transaction should be treated separately. When a trade transaction is delivered via offline processes, but subsequent follow-up transactions are delivered digitally, the follow-up transactions should be considered as digitally delivered
- A trade transaction can be delivered via multiple (digital and non-digital) modes

Online platforms play an increasingly important role in the digital economy. They facilitate economic transactions (e.g. trade in goods and services), or non-economic interactions (e.g., social media and discussion sites). In 2019, the OECD, after extensive consultations, set out a broad definition of online platforms as "a digital service that facilitates interactions between two or more distinct but interdependent sets of users (whether firms or individuals) who interact through the service via the internet" (OECD, 2019).

Digital intermediary platform (DIP) is online interfaces that facilitate, for a fee, the direct interaction between multiple buyers and multiple sellers, without the platform taking economic ownership of the goods or rendering the services that are being sold (intermediated). The assumption in this Handbook is that all transactions undertaken via a DIP are digitally ordered (OECD, WTO & IMF, 2023).

Since 2019, Bank Indonesia has begun studying the concept of digital trade in Indonesia. This is what Bank Indonesia analyse of digital economic activities in Indonesia. In its development, Bank Indonesia has formulated a concept model for digital trade transactions that occur in Indonesia (see Figure 2) which is presented as follows:



Type Transaction	Type of Order		Type of Delivery		Product	Nature
	Digital	Non	Digital	Non		
Transaction I	6			8	Goods	Digital Trade
Transaction II	4			8	Goods	Digital Trade
Transaction III	3			8	Goods	Digital Trade
Transaction IV		1	7		Services	Digital Trade
Transaction V	6		7		Services	Digital Trade
Transaction VI	4		5		Services	Digital Trade
Transaction VII	3		2		Services	Digital Trade
Transaction VIII		1		8	Goods/Services	Non Digital Trade

Figure 2. Digital Trade Transaction Model in Indonesia

In this scheme (see Figure 2) it is described into 8 (eight) digital goods and services trade transaction models. Meanwhile, transactions that use non-digital trading facilities for ordering and delivery (type 8 transactions) are not digital trading schemes.

This transaction model is the basis for forming cross-border digital trade statistics in Indonesia for both goods and services. For exports of both services and goods, it means that the customer comes from abroad (non-resident) while the supplier comes from within the country (resident). Then, for imports of both services and goods, it means that the customer comes from within the country (resident) while the supplier comes from abroad (non-resident).

2.2. Goods Export Processing

In compiling digital trade statistics on exports and imports of goods and services, Bank Indonesia, especially the Department of Statistics, has made extraordinary efforts in exploring potential data sources, data collection and data processing. The data sources that form the basis for compiling cross-border digital trade statistics are the International Transaction Reporting System (ITRS), Customs Documents, domestic e-commerce, websites/internet, and Credit and Debit Cards (LBUT). In the process of identifying digital trade actors in goods exports, the data sources used are ITRS, export platforms, as well as export transaction data from the e-commerce companies that have collaborated with Bank Indonesia. ITRS and export platforms are used as sources to obtain the names of companies indicated to be carrying out digital trade exports, which are then obtained from the transaction value of special export customs documents (export declaration). Next, selected actors will be invited to a Focus Group Discussion (FGD).

Specifically, digital trade export players are grouped into 3 (three) categories based on the ordering media, namely through DIP, internal systems and websites. Specifically, orders made via the internal system and website are obtained from ITRS. Meanwhile, orders via DIP are obtained from websites/internet, export platforms, as well as the e-commerce sites that have collaborated with Bank Indonesia.

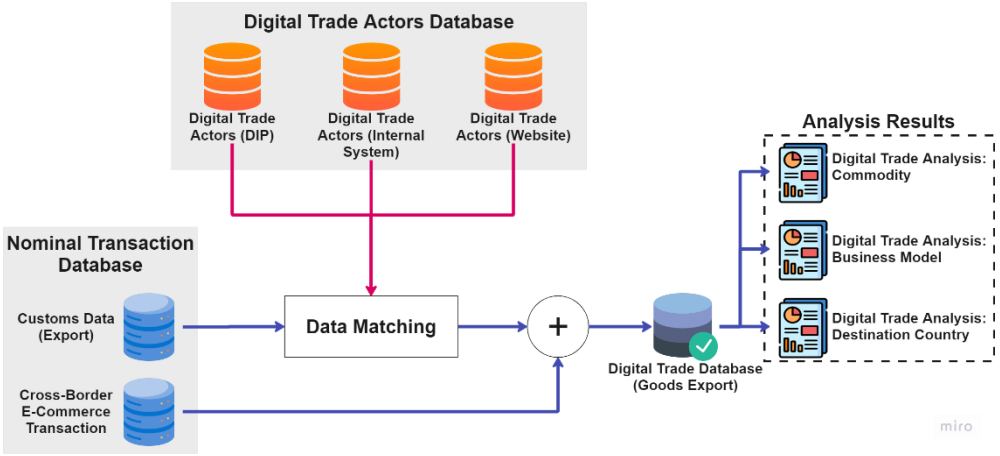


Figure 3. General Description of Goods Export Data Processing in Indonesia

Even though there are 5 (five) data sources used in compiling digital trade statistics on exports and imports of goods and services, this research only focuses on the process of identifying digital actors who export goods sourced from export platforms using the web scraping method.

2.3. Web Scraping

Web scraping is the process of extracting data off the web programmatically and transforming it into a structured dataset. Web scraping allows for larger amounts of data to be collected in a shorter span of time and in an automated fashion that minimizes errors. There are two types of web scraping. The first is screen scraping, where you extract data from source code of a website, with an HTML parser or regular expression matching. The second is using application programming interfaces, commonly referred to as APIs. This is where a website offers a set of structured HTTP requests that return JSON or XML files (Rundel & Dogucu, 2020).

Many national statistical agencies started relying on web scraping as a form of data collection, including the Italian National Institute of Statistics, ISTAT (Polidoro et al., 2015), the Federal Statistical Office of Germany, Destatis (Destatis, 2018), and Statistics Netherlands (Ten Bosch et al., 2018). One widespread way such agencies use web scraping is in automating the collection of prices of specific consumer products (e.g., electronics, housing, and medicine) to calculate some form of index of consumer prices. Uses of web scraping for data collection for other purposes have also been considered. The United States Census Bureau is building a tool that automatically scrapes tax revenue collections from websites of state and local governments as opposed to collecting this information with a traditional questionnaire (Dumbacher and Capps, 2016). Similarly, Statistics Canada (2019) is looking into ways how they can incorporate web scraping to reduce the burden on survey responders.

In industry, perhaps the best-known scraper is Googlebot (Google, 2019) which scrapes data from many web pages for Google's search engine. Web scraping is also often used in e-commerce. For instance, flight comparison websites scrape data from multiple airlines (Poggi et al., 2007). Many e-commerce websites scrape pricing information from their competitors' websites (Stiving, 2017).

Web scraping comprises three primary and intertwined phases: 1) website analysis, 2) website crawling, and 3) data organization (see Figure 4) (Krotov & Silva, 2018; Krotov & Tennyson, 2018). Each phase requires one to understand several Web technologies and at least one popular programming language, such as R or Python. However, these three phases often require at least some human involvement and, thus, one cannot fully automate them.

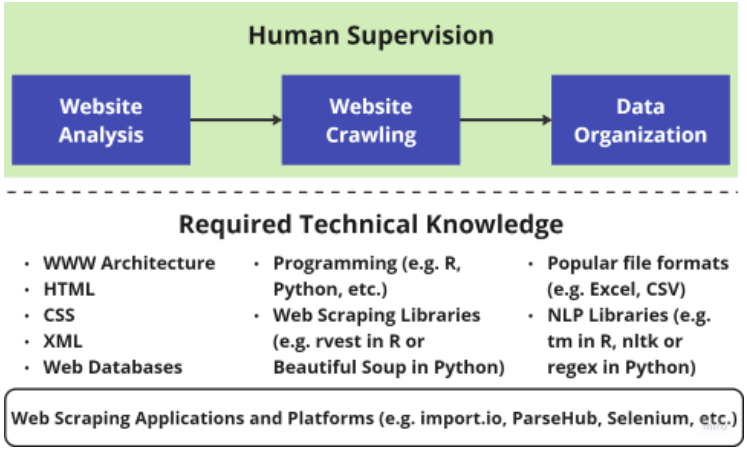


Figure 4. Web Scraping (source: Krotov & Silva, 2018; Krotov & Tennyson, 2018)

From a legal perspective, if an organization or person wants to use web scraping, they need to answer the question, namely whether their scraping action is detrimental to the website being scraped. If the scraping activity is so intense that it interferes with the services of the scraped website or the scraped data is used to duplicate the website's activities or services, then even if there is no regulation, the website has an excuse. to file a lawsuit against the scrapper. Meanwhile, from an ethical point of view, considering that web scraping already has many use cases and professional providers in the market, we can claim that there is no harm in using web scraping for business purposes. (Krotov et al., 2020).

In terms of using web scraping to identify digital trade actors, we do this to retrieve the names of exporters who are suppliers on the export platforms. The data provided from this platform is open to the public and the period used for scraping is once a month so that it does not cause problems in terms of legality or in other words the process carried out is an act that does not violate the law.

3. Methodology

3.1. Data Source

In compiling cross-border digital trade statistics, especially exports of goods, 5 (five) data sources are used as a basis, namely:

1. International Transaction Reporting System (ITRS)

ITRS is data obtained from bank reporting to the Central Bank regarding transactions in the context of international trade. This database is designed to track

and report international trade activities including exports and imports. ITRS data is the basis for identifying actors who export goods and services.

2. Export Platforms

In order to develop the potential of the export business sector in Indonesia, there are digital export platforms that can be used to help domestic export business actors to gain access and information regarding the needs of prospective buyers abroad. Two of these export platforms are used as web scraping objects to identify digital trade actors in exporting goods. These platforms provide information about export opportunities as well as connect Indonesian suppliers with buyers around the world and promote Indonesian companies and products online to reach a wider range of potential buyers.

3. Export Declaration (Customs Data in Export)

Export data includes all goods brought outside Indonesian territory. This data was obtained from the Directorate General of Customs and Excise (under the Indonesian Ministry of Finance) which has collaborated with Bank Indonesia. Export data is the basis for obtaining transaction values from actors identified in points 1 and 2 as digital trade actors.

4. Focus Group Discussion (FGD)

FGD with the company was carried out to deepen the company's business processes, especially the practice of ordering digitally and delivering digitally. Representative actors are selected and sorted based on their transaction value and then a bilateral structured discussion is held with Bank Indonesia. Open questions were asked to collect information regarding the extent to which the company has implemented digital technology in the process of selling the goods it markets, because the process of ordering and sending goods is the main key in identifying digital trading actors. The results of the FGD were used as the basis for analysis carried out by Bank Indonesia to determine whether the company was a digital trading actor or not. If the company does not carry out digital trade practices, export transactions will be excluded from the digital trade statistics for exports of goods (filtering process).

5. Cross Border E-commerce Transaction Data

Goods export data used as an additional data source besides customs data is export transaction data originating from domestic e-commerce companies that have collaborated with Bank Indonesia.

In general, the description of the mapping of the data universe used in the process of compiling digital trade statistics on exports of goods is as follows (see Figure 5):

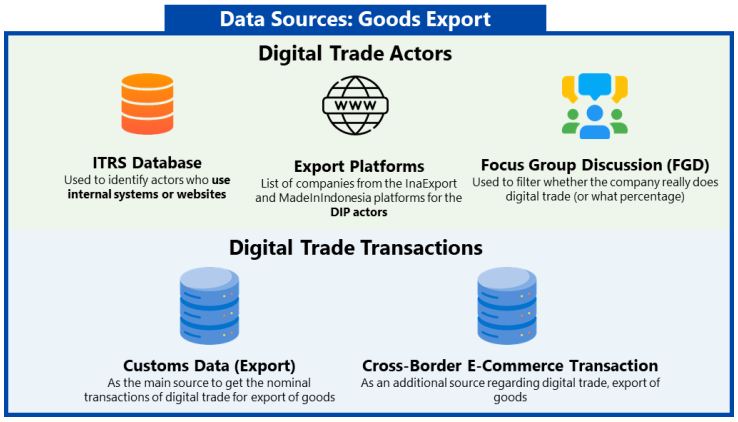


Figure 5. Data source mapping for Goods Export

3.2. Data Processing

In compiling cross-border digital trade statistics, there are 2 (two) major processes, namely (i) identification of cross-border digital trade actors and (ii) matching of actors with the nominal transactions. These two things are done sequentially, starting from the first process (identifying the digital trade actors) then matching the transaction nominal. For more details, each process is as follows:

1) Identify cross-border digital trade actors

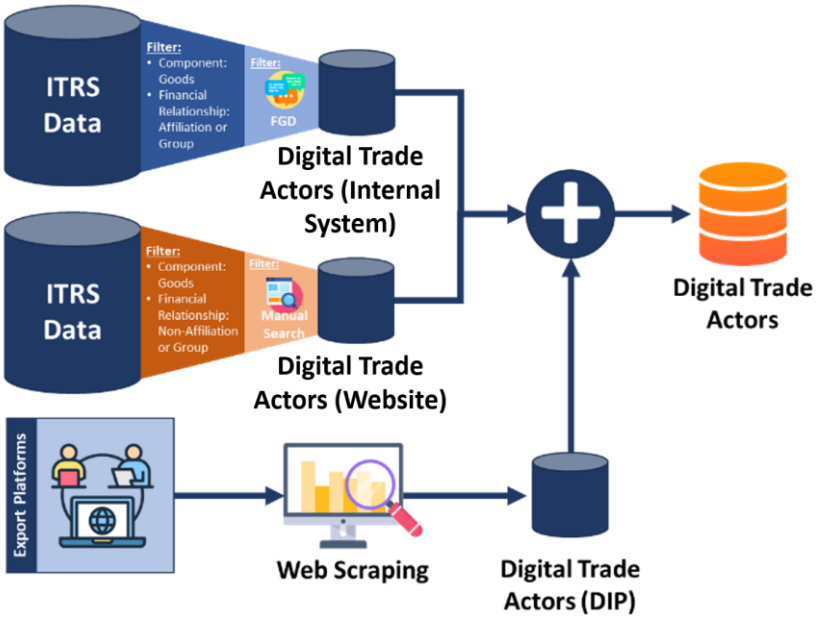


Figure 6. Digital Trade Actor Identification Process (Export of Goods)

Based on (Figure 6), after obtaining various companies exporting goods from ITRS data, the process of identifying these companies one by one via the internet was carried out regarding the ordering media used both through internal systems, websites and DIP. A company is categorized as ordering goods through an internal system if the transaction is carried out between the parent company and a subsidiary (having a group/affiliate financial relationship) and has

been confirmed through an FGD. Meanwhile, a company is categorized as ordering goods via a website if the company has a website that is used to market and order products. In addition, the actors through the Digital Intermediary Platform (DIP) were identified using a web scraping process by extracting the names of supplier/seller companies from existing export platforms in Indonesia.

2) Identify the nominal value of cross-border digital trade transactions

After the process of identifying the digital trade actors and the ordering media used, the next step is to match the name of the digital trade actors with special customs data for Exported Goods (export declaration) (see Figure 7). The data matching process uses a gestalt pattern matching algorithm (one of the algorithms in the entity resolution family that is part of advanced analytics).

If the export transaction value has been obtained from export declaration, it is necessary to carry out special screening of transaction actors via the website. Conformity between products marketed on the website and export commodities recorded in the PEB database is very important to be checked to increase data accuracy.

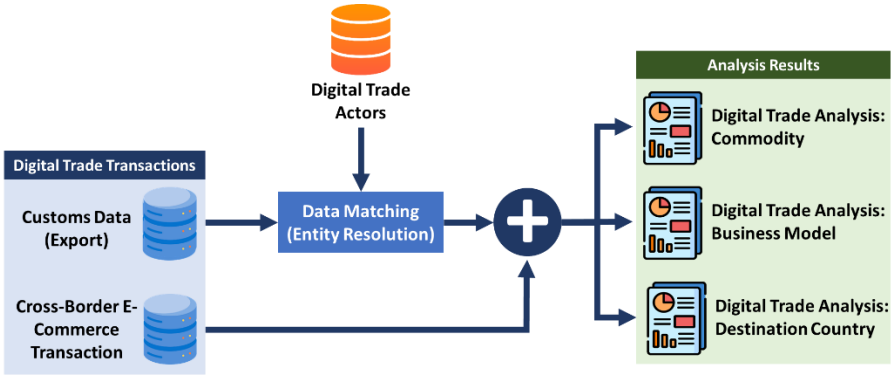


Figure 7. Cross-Border Digital Trade Transaction Nominal Identification Process

In the end, the value of goods export transactions resulting from data matching and the value of goods export transactions originating from 2 (two) e-commerce sites are used to compile digital trade statistics for goods exports.

4. Result and Discussion

In compiling digital goods export trade statistics, the first thing to do is carry out web scraping to obtain a list of digital trade actors through DIP channels, while the website and internal systems come from a list of actors collected manually. Web scraping is done to retrieve data from a table or list on the export platforms. The steps taken in web scraping are (Figure 8):

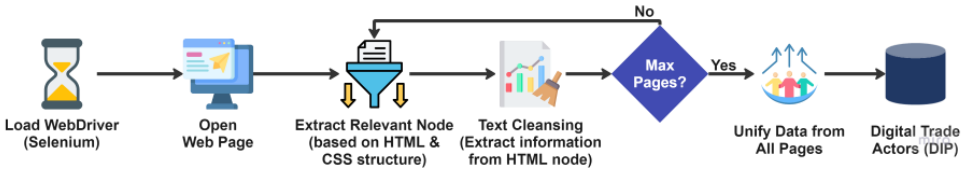


Figure 8. Web scraping steps

In the web scraping process, the information we want to obtain from the export platforms is the name of the company that are determined to be a supplier of the platform.

After obtaining the names of suppliers from all export platforms, the list of actors will be combined with other channels. The next process carried out is to match the name of the digital trade actor who exports goods for all routes with the name of the export actor in the transaction database (export declaration) so that the nominal export transaction carried out by that actor is obtained. This matching process uses advanced analytical techniques, namely entity resolution using a gestalt pattern matching algorithm which is formulated as follows:

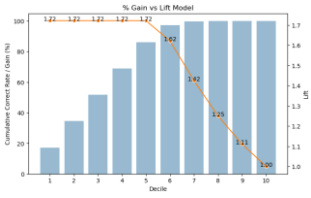
$$D = \frac{2K_m}{|S_1| + |S_2|}$$

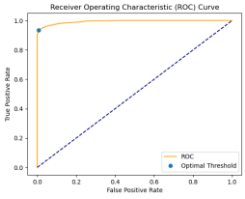
Notes:

- K_m is the number of matching characters
- S_1 is the first string
- S_2 is the second string

Using this algorithm, an analysis of the best threshold is carried out that provides the highest accuracy using decile analysis and receiver operating characteristic (ROC) curve analysis as shown in (Figure 9):

decile	decile_size	min_proba	max_proba	decile (%)	cum_non_takers_rate	cum_takers_rate / gain	ks	lift
1	750	1	1	10%	0%	17%	0.172176	1.721763
2	750	1	1	20%	0%	34%	0.344353	1.721763
3	750	1	1	30%	0%	52%	0.516529	1.721763
4	750	1	1	40%	0%	69%	0.688705	1.721763
5	750	1	1	50%	0%	86%	0.860882	1.721763
6	750	0.848484848	1	60%	8%	97%	0.889578	1.621518
7	750	0.774193548	0.848484848	70%	29%	100%	0.709077	1.424636
8	750	0.714285714	0.774193548	80%	52%	100%	0.477099	1.25
9	750	0.647058824	0.714285714	90%	76%	100%	0.23855	1.111111
10	750	0.3125	0.647058824	100%	100%	100%	0	1





(a) (b)

Figure 9. Threshold analysis using (a) Decile analysis (b) ROC curve

Based on the analysis results in (Figure 9) it is known that: the best decile used is the 6th decile which has a true positive rate of 97% with a lift showing 1.62 or in other words this algorithm has a reliability of 1.62 times compared to using the similarity threshold. of 0.5, by using the ROC curve the optimal threshold was also obtained, namely 0.9268 with an accuracy of 96% by utilizing 7500 data as testing.

By using advanced analytics, namely web scraping and entity resolution as method development, several benefits are obtained, namely:

a) The scope of digital trade actors

The application of advanced analytics in terms of compiling statistics, especially in terms of data collection, has had a major impact in increasing the scope of identified digital trade actors, especially those using DIP

channels. The additional actor coverage is 294 actors (before the use of web scraping there were only 153 actors, but after there are 447 actors that use DIP as a transaction intermediary). This shows that there has been an increase in actors by 292% with the implementation of web scraping in identifying actors of cross-border exports of digital trade goods. Subsequently, the expansion of identified actors will lead to a greater transaction value.

b) The efficiency of the work process

The utilization of advanced analytics contributes to enhancing the efficiency of work processes at Bank Indonesia, where statistical compilation results become available more quickly, even with limited resources. The efficiency impact of implementing advanced analytics is presented as follows (see Figure 10):

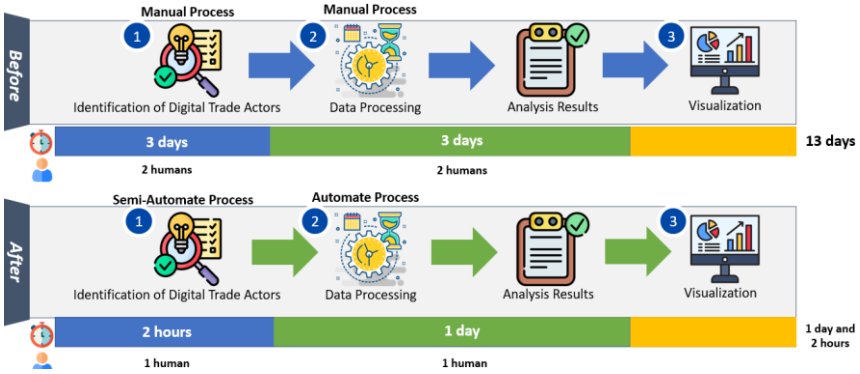


Figure 10. Business Process Improvement (effect of Advanced Analytics)

Based on (Figure 10), there is efficiency in the work process so that there is a reduction in the time required for a work process (13 days to 1 day 2 hours) thereby reducing the work process to around 12 days. In addition, in terms of human resource requirements, it is found that there is a reduction in resources. What is needed is from 2 people to 1 person. This is due to the use of programming in it so that it no longer requires human touch starting from the identification and data processing process.

c) The quality of the data obtained

Based on improvements in terms of work processes and the web scraping process itself, apart from being able to increase the scope of actors, it also increases the effectiveness and efficiency of the work process. The quality of data obtained from web scraping results is also increasing, this is because the web scraping process uses a programming method so that there is no human touch in it, thereby minimizing errors caused in data processing (Rundel & Dogucu, 2020).

5. Conclusion and Future Works

Based on this discussion, several conclusions were obtained, namely:

a) By implementing advanced analytics in the process of compiling cross-border digital trade statistics, especially the use of web scraping, it can increase the scope of exporters of goods by 294 actors (an increase of 292%).

b) By implementing advanced analytics (a combination of web scraping and entity resolution) work process efficiency can be achieved, namely by reducing the work process by around 12 days, in terms of human resources required can be reduced from initially requiring 2 people to only 1 person.

c) By using web scraping in the process of compiling statistics, the quality of the data used can be guaranteed, or in other words, without human touch, human error can be minimized.

In addition, the web scraping process is currently limited to DIP-mediated transactions only. In the future, Bank Indonesia will develop this web scraping technique to identify actors who use website intermediaries (based on whether the website can be used for transactions or not) and develop web scraping to extract the profile of a company so that it can be tracked (used as a reference for knowledge, whether the company is a digital trading actor that uses an internal system or not). This is done considering that the initial data known as the basis for identifying digital trade actors is only based on the name of the company, so it is necessary to extract new basic data sources to identify digital trade actors.

References

Achmad, N., & Schreyer, P. (2016). Measuring GDP in a Digitalised Economy. *OECD Statistics Working Papers*. Retrieved from https://doi.org/10.1787/5jlwqd81d09r-en

Alisjahbana, A., Setiawan, M., Effendi, N., Santoso, T., & Hadibrata, B. (2020). The Adoption of Digital Technology and Labor Demand in the Indonesian Banking Sector. *International Journal of Social Economics, 47*(9), 1109-1122.

Destatis. (2018). *"Confidentiality in the 2021 Census", Methods – Approaches – Developments: Information of the German Federal Statistical Office*. Wiesbaden.

Dogucu, M., & Çetinkaya-Rundel, M. (2021). Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities. *Journal of Statistics and Data Science Education*, S112-S122. doi:10.1080/10691898.2020.1787116

Dumbacher, B., & Capps, C. (2016). Big Data Methods for Scraping Government Tax Revenue From the Web. *Proceedings of the Joint Statistical Meetings, Section on Statistical Learning and Data Science*, 2940–2954.

Ezel, S., & Koester, S. (2023). *Transforming Global Trade and Development With Digital Technologies*. Information Technology & Innovation Foundation.

Google. (2019). *Googlebot*. Retrieved from https://support.google.com/webmasters/answer/182072?hl=en

Google, Temasek, and Bain & Company. (2022). *Through the Waves Towards a Sea of Opportunity*. Retrieved from Google e-Conomy SEA: https://economysea.withgoogle.com/home/

IMF. (2023). *World Economic Outlook: Navigating Global Divergences*. Retrieved from IMF: https://www.imf.org/en/Publications/WEO/Issue/2023/10/10/world-economic-outlook-october-2023

Indonesia's Central Agency of Statistics (BPS). (2023). *Indonesia's Statistics*. Jakarta: Indonesia's Central Agency of Statistics (BPS).

Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N., & Roig-Tierno, N. (2021). Digital Transformation: An Overview of the Current State of the Art of Research. *SAGE Open, 11*(3).

Krotov, V., & Silva, L. (2018). Legality and ethics of Web scraping. *Proceedings of the 24th Americas Conference on Information Systems*.

Krotov, V., & Tennyson, M. (2018). Scraping financial data from the Web using the R language. *Journal of Emerging Technologies in Accounting, 15*(1), 169-181.

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems, 47*, 539-563. doi:10.17705/1CAIS.04724

OECD. (2019). *An Introduction to Online Platforms and Their Role in the Digital Transformation*. Retrieved from https://doi.org/10.1787/53e5f593-en

OECD and IMF. (2017). *Measuring Digital Trade: Results of OECD/IMF Stocktaking Survey*. Thirtieth Meeting of the IMF Committee on Balance of Payments Statistics.

OECD, WTO, & IMF. (2023). *Handbook on Measuring Digital Trade*.

Poggi, N., Berral, J., Moreno, T., Gavalda, R., & Torres, J. (2007). Automatic Detection and Banning of Content Stealing Bots for e-Commerce. *NIPS 2007 Workshop on Machine Learning in Adversarial Environments for Computer Security, 2*.

Polidoro, F., Giannini, R., Conte, R., Mosca, S., & Rossetti, F. (2015). Web Scraping Techniques to Collect Data on Consumer Electronics and Airfares for Italian HICP Compilation. *Statistical Journal of the IAOS, 31*, 165–176. doi:10.3233/sji-150901

Sapulette, M., & Muchtar, P. (2023). Redefining Indonesia's Digital Economiy. *Economic Research Institute for ASEAN and East Asia*.

Sapulette, M., & Santoso, T. (2021). Macroeconomic and Public Health Policies amid COVID-19 Pandemic: Global FInancial Sector's Reponses. *5*(2), 91-102.

Statistics Canada. (2019). *Web Scraping*. Retrieved from https://www.statcan.gc.ca/eng/our-data/where/web-scraping

Stiving, M. (2017). B2b Pricing Systems: Proving ROI. In A. Hinterhuber, & S. Liouzu, *Innovation in Pricing* (pp. 137–144). London: Routledge.

Suominen, A., & Hajikhani, A. (2021). Research themes in big data analytics for policymaking: Insights from a mixed-methods systematic literature review. *Policy Internet, 13*, 464-484. doi:10.1002/poi3.258

Ten Bosch, O., Windmeijer, D., van Delden, A., & van den Heuvel, G. (2018). Web Scraping Meets Survey Design: Combining Forces. *Big Data Meets Survey Science Conference*.

Ventures, E. (2022). *Digital Competitiveness Index 2022: Towards Indonesia's Digital Golden Era*. Retrieved from https://east.vc/reports/east-ventures-digitalcompetitiveness-index-2022/

BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

DSta
Departemen Statistik
Bank Indonesia

Web Scraping as a Detection Tool in Identifying Indonesian Cross-Border Digital Trade Actors

Aditya Wisnugraha Sugiyarto, Hasudungan Paulanka Siburian, Marlina Novita Uligoma, Dwi Cahyo Ardianto, Novi Ajeng Salehah, Detasya Avri Magfira

Department of Statistics, 13th February 2024

BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

Motivation and Introduction

Indonesia: Rising Digital Frontier

Volume
42% of the ASEAN Digital Economy

Investment Inflows
100% Growth from 2021 to 2022
from $4.5 bio to $9.1 bio

Internet User
66% Of whole population in Indonesia

Population
52% Of whole population dominated by gen Y and Z

Digital Economy Growth
8x In just less than 10 years

Source: Google, World Bank, NSO, IMF, Sapulette & Santoso, etc. (2022)

Current State

- Identification of digital trade actors process becomes increasingly complex and time consuming
- Available information is often limited to company names
- The current database has incomplete transaction coverage due to the manual data processing

Opportunity

- The importance of data in supporting policy and decision making has encouraged the development of big data analysis (Suominen & Hajikhani, 2021)
- The use of big data analytics provides the ability to analyze and interpret large and complex amounts of data
- Web scraping can become a very effective tool for accessing information about digital trade actors, including their activities and characteristics (Rundel & Dogucu, 2020)
- Web scraping can be used to gather information from various digital trading platforms, company directories, and other related sources (Rundel & Dogucu, 2020)

Desired State

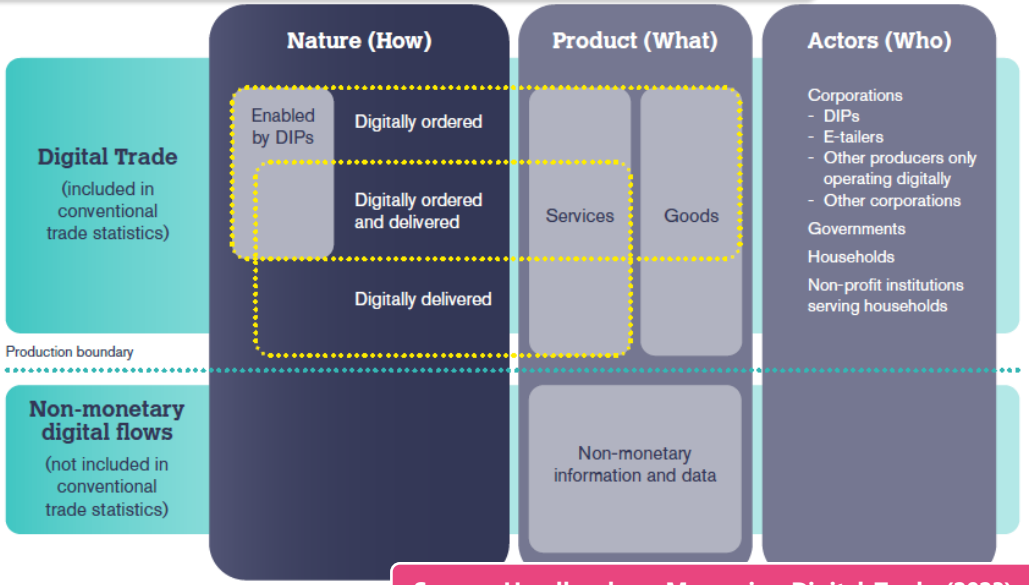
- The identification process of digital trade actors is clear and efficient
- The information available from digital trade actors is complete
- A new database that has very good quality with complete transaction coverage

Paper Contribution

- This paper contributes in terms of providing one of the first work/paper related to cross-border digital trade in Indonesia.
- In addition, the authors provide overview and framework of how to use advanced analytics to improve the quality of digital trade statistics compilations

Digital Trade Concept

Conceptual Framework of Digital Trade



Source: Handbook on Measuring Digital Trade (2023)

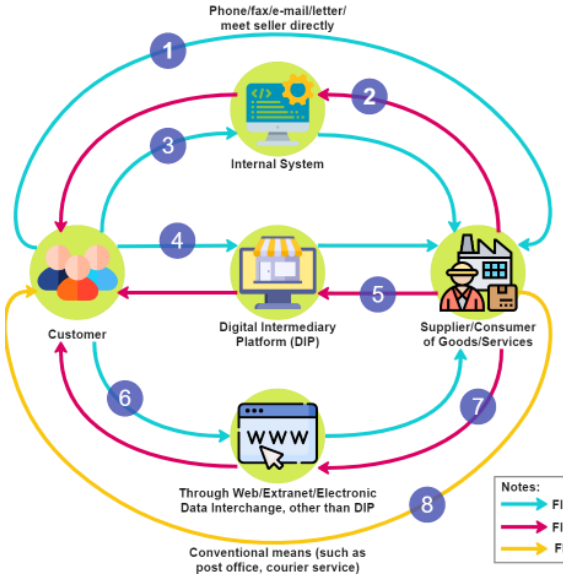
Based on Conceptual Framework of Digital Trade, digital trade is divided into two types of transactions (how), namely:

- **Transactions for goods/services ordered digitally (digitally ordered)**
  Cross-border digital transactions ordered digitally are the international sale or purchase of a good or service, conducted over computer networks by methods specifically designed for the purpose of receiving or placing orders.
- **Transactions for goods/services sent digitally (digitally delivered)**
  Digitally delivered trade is defined as all international trade transactions that are delivered remotely over computer networks

Since 2019, Bank Indonesia has begun studying the concept of digital trade in Indonesia.

In its development, **Bank Indonesia has formulated a concept model for digital trade transactions that occur in Indonesia** (see Digital Trade Transaction Model in Indonesia).

Digital Trade Transaction Model in Indonesia

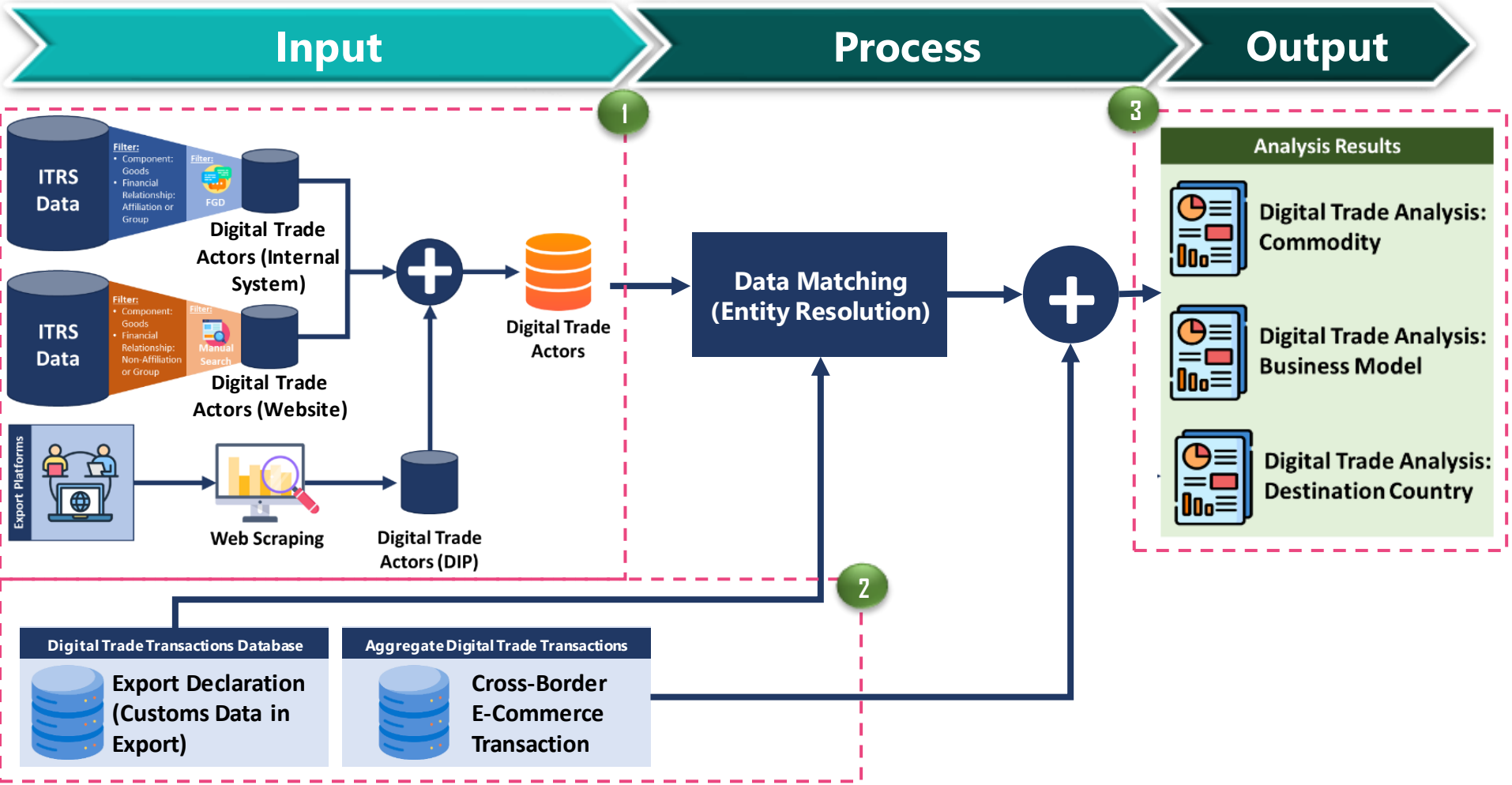


Type Transaction	Type of Order		Type of Delivery		Product	Nature
	Digital	Non	Digital	Non		
Transaction I	6			8	Goods	Digital Trade
Transaction II	4			8	Goods	Digital Trade
Transaction III	3			8	Goods	Digital Trade
Transaction IV		1	7		Services	Digital Trade
Transaction V	6		7		Services	Digital Trade
Transaction VI	4		5		Services	Digital Trade
Transaction VII	3		2		Services	Digital Trade
Transaction VIII		1		8	Goods/Services	Non Digital Trade

This transaction model is the basis for forming cross-border digital trade statistics in Indonesia for both goods and services.

Methodology

Compiling cross-border digital trade statistics on goods exports involves 3 main processes: profiling of digital trade actors, identification of transactions, and extracting analysis results of the statistics...



1. Identification (profiling) of cross-border digital trade actors on goods export using Digital Intermediary Platform (DIP), websites, internal systems channels by utilizing data sources such as ITRS and export platforms.
2. Based on data from digital trade actors, goods export transactions are identified based on export declarations (customs data on exports) and supplemented with aggregate data on cross-border e-commerce transactions. This matching process uses the entity resolution method.
3. After obtaining information related to transactions from digital trade actors, statistical results and details of cross-border digital trade are obtained.

Results

1 Identification of Digital Trade Actors

- The coverage of actors has increased by 292%.
- The use of web scraping improve work process efficiency.
- The expansion of identified actors will lead to a greater transaction coverage.

2 Data Processing

- The entity resolution method (gestalt pattern matching) with 96% accuracy can increase the efficiency of work process in terms of processing time and human resources.

3 Analysis Results of The Statistics

- **A more detailed database of cross-border digital trade statistics is acquired**, such as names of actors, transaction values, commodities, destination countries, where these information can be interlinked to obtain a more in-depth analysis.
- **The new database has better quality compared to the manual process, featuring more accurate actor names and broader transaction coverage**.

The Efficiency of the work process

	Before	After
Identification of Digital Trade Actors	3 days, 2 persons	2 hours, 1 person
Data Processing	10 days, 2 persons	1 day, 1 person
TOTAL	13 days, 2 persons	1 day and 2 hours, 1 person

Conclusion & Way Forward

Conclusion

- Advanced analytics (web scraping and entity resolution) can increase the scope of exporters of goods by 294 actors (an increase of 292%).
- Efficiency in the work process has increased as reflected in faster data processing (13 days to 1 day) and fewer human resources (2 humans to 1 human).
- The quality of data can be enhanced through advanced analytics as it can minimize the human error.

Way Forward

- Utilize web scraping to identify actors that use website intermediaries (based on whether the website can be used for transactions or not), as one of the data sources for measuring cross-border digital trade in Indonesia.
- Develop web scraping to extract the profile of companies so that it can be tracked (as a reference in getting information whether the company is a digital trade actor that use an internal system or not).

Bibliography

Achmad, N., & Schreyer, P. (2016). Measuring GDP in a Digitalised Economy. *OECD Statistics Working Papers*. Retrieved from https://doi.org/10.1787/5jlwqd81d09r-en

Alisjahbana, A., Setiawan, M., Effendi, N., Santoso, T., & Hadibrata, B. (2020). The Adoption of Digital Technology and Labor Demand in the Indonesian Banking Sector. *International Journal of Social Economics, 47*(9), 1109-1122.

Destatis. (2018). *"Confidentiality in the 2021 Census", Methods – Approaches – Developments: Information of the German Federal Statistical Office*. Wiesbaden.

Dogucu, M., & Çetinkaya-Rundel, M. (2021). Web Scraping in the Statistics and Data Science Curriculum: Challenges and Opportunities. *Journal of Statistics and Data Science Education*, S112-S122. doi:10.1080/10691898.2020.1787116

Dumbacher, B., & Capps, C. (2016). Big Data Methods for Scraping Government Tax Revenue From the Web. *Proceedings of the Joint Statistical Meetings, Section on Statistical Learning and Data Science*, 2940–2954.

Ezel, S., & Koester, S. (2023). *Transforming Global Trade and Development With Digital Technologies*. Information Technology & Innovation Foundation.

Google. (2019). *Googlebot*. Retrieved from https://support.google.com/webmasters/answer/182072?hl=en

Google, Temasek, and Bain & Company. (2022). *Through the Waves Towards a Sea of Opportunity*. Retrieved from Google e-Conomy SEA: https://economysea.withgoogle.com/home/

IMF. (2023). *World Economic Outlook: Navigating Global Divergences*. Retrieved from IMF: https://www.imf.org/en/Publications/WEO/Issue/2023/10/10/world-economic-outlook-october-2023

Indonesia's Central Agency of Statistics (BPS). (2023). *Indonesia's Statistics*. Jakarta: Indonesia's Central Agency of Statistics (BPS).

Kraus, S., Jones, P., Kailer, N., Weinmann, A., Chaparro-Banegas, N., & Roig-Tierno, N. (2021). Digital Transformation: An Overview of the Current State of the Art of Research. *SAGE Open, 11*(3).

Krotov, V., & Silva, L. (2018). Legality and ethics of Web scraping. *Proceedings of the 24th Americas Conference on Information Systems*.

Krotov, V., & Tennyson, M. (2018). Scraping financial data from the Web using the R language. *Journal of Emerging Technologies in Accounting, 15*(1), 169-181.

Krotov, V., Johnson, L., & Silva, L. (2020). Tutorial: Legality and Ethics of Web Scraping. *Communications of the Association for Information Systems, 47*, 539-563. doi:10.17705/1CAIS.04724

OECD. (2019). *An Introduction to Online Platforms and Their Role in the Digital Transformation*. Retrieved from https://doi.org/10.1787/53e5f593-en

OECD and IMF. (2017). *Measuring Digital Trade: Results of OECD/IMF Stocktaking Survey*. Thirtieth Meeting of the IMF Committee on Balance of Payments Statistics.

OECD, WTO, & IMF. (2023). *Handbook on Measuring Digital Trade*.

Poggi, N., Berral, J., Moreno, T., Gavalda, R., & Torres, J. (2007). Automatic Detection and Banning of Content Stealing Bots for e-Commerce. *NIPS 2007 Workshop on Machine Learning in Adversarial Environments for Computer Security*, *2*.

Polidoro, F., Giannini, R., Conte, R., Mosca, S., & Rossetti, F. (2015). Web Scraping Techniques to Collect Data on Consumer Electronics and Airfares for Italian HICP Compilation. *Statistical Journal of the IAOS, 31*, 165–176. doi:10.3233/sji-150901

Sapulette, M., & Muchtar, P. (2023). Redefining Indonesia's Digital Economiy. *Economic Research Institute for ASEAN and East Asia*.

Sapulette, M., & Santoso, T. (2021). Macroeconomic and Public Health Policies amid COVID-19 Pandemic: Global FInancial Sector's Reponses. *5*(2), 91-102.

Statistics Canada. (2019). *Web Scraping*. Retrieved from https://www.statcan.gc.ca/eng/our-data/where/web-scraping

Stiving, M. (2017). B2b Pricing Systems: Proving ROI. In A. Hinterhuber, & S. Liouzu, *Innovation in Pricing* (pp. 137–144). London: Routledge.

Suominen, A., & Hajikhani, A. (2021). Research themes in big data analytics for policymaking: Insights from a mixed-methods systematic literature review. *Policy Internet, 13*, 464-484. doi:10.1002/poi3.258

Ten Bosch, O., Windmeijer, D., van Delden, A., & van den Heuvel, G. (2018). Web Scraping Meets Survey Design: Combining Forces. *Big Data Meets Survey Science Conference*.

Ventures, E. (2022). *Digital Competitiveness Index 2022: Towards Indonesia's Digital Golden Era*. Retrieved from https://east.vc/reports/east-ventures-digitalcompetitiveness-index-2022/

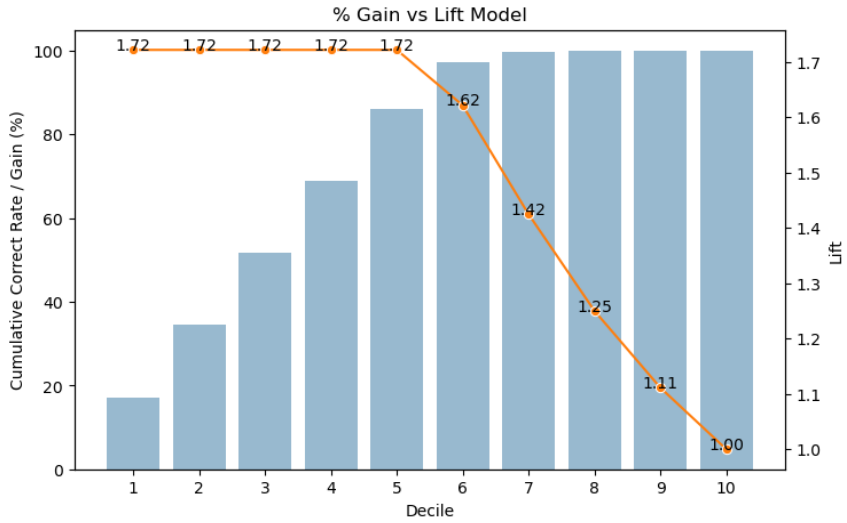
BANK INDONESIA
BANK SENTRAL REPUBLIK INDONESIA

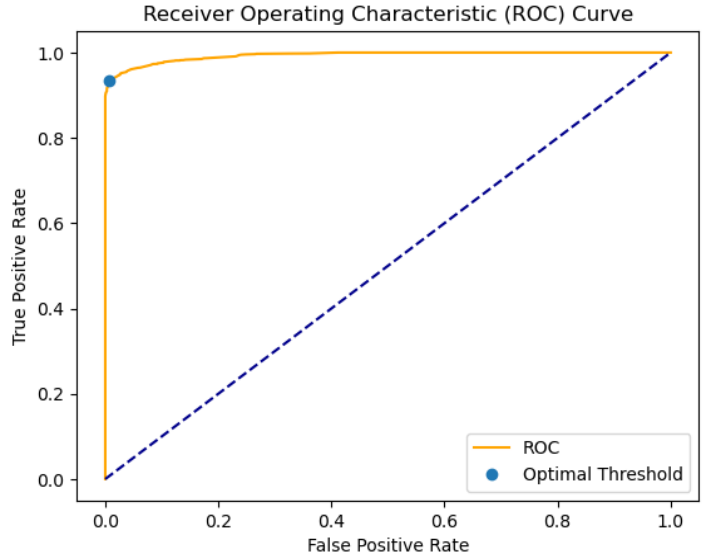
THANK YOU

Entity Resolution: Accuracy

Decile Analysis

decile	decile_size	min_proba	max_proba	decile (%)	cum_non_takers_rate	cum_takers_rate / gain	ks	lift
1	750	1	1	10%	0%	17%	0.172176	1.721763
2	750	1	1	20%	0%	34%	0.344353	1.721763
3	750	1	1	30%	0%	52%	0.516529	1.721763
4	750	1	1	40%	0%	69%	0.688705	1.721763
5	750	1	1	50%	0%	86%	0.860882	1.721763
6	750	0.848484848	1	60%	8%	97%	0.889578	1.621518
7	750	0.774193548	0.848484848	70%	29%	100%	0.709077	1.424636
8	750	0.714285714	0.774193548	80%	52%	100%	0.477099	1.25
9	750	0.647058824	0.714285714	90%	76%	100%	0.23855	1.111111
10	750	0.3125	0.647058824	100%	100%	100%	0	1





Based on the results of the analysis using Decile Analysis and ROC-AUC to obtain the best threshold results, it was found that::

- **The best decile obtained was the 6th decile, which obtained an improvement of 97%** (a true positive rate of around 97% was obtained in that decile) and the improvement still showed 1.62 (this model is 1.62x better than using the default threshold, 0.5).
- By using the AUC-ROC curve**, the optimal threshold value was obtained is 0.9268 with an accuracy of 96%** based on a dataset of 7500.