QUBAHAN ECOLEAD JOURNAL
VOL. 4, NO. 1, March 2025
https://doi.org/10.48161/qaj.v4n1a22

Market Composition and Data-Quality Constraints in Global Listed-Securities Reference Files: Evidence from a Pre-2020 Cross-Market Dataset

Rania Lampou[1], Abdul Malik Iskandar[2], Nozim Kurbonov[3]
[1]Directorate of Educational Technology and Innovation, Greek Ministry of Education, Athens, Greece
[2]Faculty of Teaching & Education Sciences, Megarezky University, Makassar, Indonesia;
[3]Tashkent University of Information Technologies named after Muhammad Al-Khwarizmi,Tashkent, Uzbekistan
Corresponding author email: rania.lampou@gmail.com;

ABSTRACT: This study examines the empirical usability of pre-2020 global listed securities reference data by assessing market composition, venue concentration, and identifier completeness. Using a source file of 237 raw records and 56 fields, the analysis retains 147 observations created before 2020 and applies descriptive frequency analysis, cross-tabulation, robust summary statistics, and field-completeness checks. The findings show that the retained sample is dominated by exchange-traded funds and common equity, which together account for nearly the entire dataset. Listings are geographically diverse but concentrated in a limited number of countries, particularly GB, AU, HK, US, and NL, while trading activity is similarly clustered across major venues and currencies. Security-level identifiers, including ISIN, SEDOL, issuer name, security description, and trading currency, are complete or near-complete, supporting reproducible descriptive and cross-market analysis. However, materially weaker LEI and GICS coverage limits the dataset's suitability for issuer-level linkage and sector-based empirical designs without supplementary enrichment. The study contributes by reframing securities reference data as research infrastructure rather than passive administrative metadata and by proposing a Reference-Data Usability Assessment Framework (RDUAF) to guide pre-analysis evaluation of dataset coverage, completeness, and research suitability. Overall, the evidence suggests that the examined file is well-suited to security-level market-structure profiling prior to the COVID-19 period but requires additional data integration to support more advanced firm-level, sectoral, or causal finance research.

Keywords: Listed Securities; Reference Data; Market Structure; Identifier Completeness; Exchange-Traded Funds; Empirical Finance.

I. INTRODUCTION

Reference data are a foundational yet underexamined layer of empirical research in capital markets. By linking securities to issuers, exchanges, countries, currencies, and classification systems, they determine whether researchers can construct samples consistently, compare instruments across markets, and reproduce database linkages with confidence. Yet finance studies typically use such files as background infrastructure rather than as empirical objects whose scope and completeness materially shape inference.

That omission is especially important in a pre-2020 setting. For many cross-market studies, the years before the COVID-19 disruption provide a relatively stable baseline for describing market organization, listing patterns, and the evolution of exchange-traded products. A careful assessment of pre-2020 securities reference data can therefore clarify both what a dataset captures and which forms of downstream analysis it can credibly support.

The source file examined here is a multi-country securities master dataset containing security descriptions, issuer names, ISIN, FIGI, SEDOL, CFI, LEI, GICS, exchange codes, listing-country codes, trading currencies, lot sizes, and shares outstanding. These fields make it possible to examine market composition across countries and venues while also evaluating the quality of the identifiers required for reproducible empirical work, external linkage, and product-level classification.

This paper makes four contributions. First, it documents the composition of the retained pre-2020 sample across listing countries, exchanges, currencies, and security types. Second, it evaluates the completeness of the identifiers and classifications that condition the analytical value of the dataset. Third, it translates those descriptive patterns into implications for future cross-market research design. Fourth, it proposes a Reference-Data Usability Assessment Framework (RDUAF) that converts evidence on coverage, concentration, and completeness into a structured assessment of research usability. The remainder of the paper reviews the relevant literature, states the research questions, explains the methodology, reports the results, discusses their implications, and concludes.

II. LITERATURE REVIEW

Research published during 2010–2020 shows that market quality depends not only on asset characteristics, but also on trading technology and venue design. Hendershott, Jones, and Menkveld [3] show that algorithmic trading can improve quoted liquidity, while Menkveld [4] and Brogaard, Hendershott, and Riordan [5] document the role of high-frequency intermediaries in price discovery and contemporary market making. Budish, Cramton, and Shim [6] argue that speed competition can distort incentives under continuous limit order books, and O'Hara [7] synthesizes how high-frequency activity altered market microstructure over the period.

Related work indicates that transparency, order routing, and information arrival affect trading outcomes across venues. Indriawan [1] documents changes in market quality around macroeconomic news announcements, Brolley [2] analyzes execution risk across lit and dark markets, and Foley and Putnins [8] show that the effects of dark trading depend on market conditions and venue structure. This literature supports treating exchange codes, security types, and trading currencies as analytically meaningful variables rather than as purely administrative metadata.

The cross-listing literature emphasizes investor recognition, governance, trading location, and the global allocation of capital. Athanassakos et al. [9] examine investor demand for cross-listed firms, while Doidge, Karolyi, and Stulz [10] show how financial globalization shifted new equity issuance outside the United States. Dodd [11] reviews the main theories of cross-listing, including liquidity, visibility, legal bonding, and market-segmentation explanations.

Subsequent evidence links cross-listing to information production, governance, and venue choice. Li, Brockman, and Zurbruegg [12] show that cross-listing interacts with firm-specific information and corporate governance, Sarkissian and Schill [13] document wave-like patterns in cross-listing activity, and Cumming, Hou, and Wu [14] show that exchange rules and governance influence the trading location of cross-listed shares. These studies justify attention to listing-country codes, exchange identifiers, and trading currencies when interpreting the geographic footprint of the sample.

Exchange-traded funds are central to the present study because ETFs comprise the largest share of the retained sample. IOSCO [15] and the U.S. Securities and Exchange Commission [16] describe ETFs as products with distinct primary- and secondary-market mechanisms, emphasizing creation-redemption processes, arbitrage, disclosure, and resilience. Madhavan and Sobczyk [17] explain ETF price dynamics and liquidity, while Israeli, Lee, and Sridharan [18] identify information-related effects associated with ETF trading and ownership.

Additional evidence shows that ETF activity can influence pricing efficiency, correlations, volatility, and short-run demand conditions. Petajisto [19] documents persistent ETF pricing inefficiencies, Da and Shive [20] link ETFs to asset-return comovement, Ben-David, Franzoni, and Moussawi [21] study ETF-related volatility effects, Broman and Shum [22] examine flows and relative liquidity, Box, Davis, and Fuller [23] analyze ETF competition and market quality, and Laipply and Madhavan [24] study fixed-income ETF

pricing and liquidity during the 2020 crisis. Accordingly, separating ETFs from ordinary shares is essential for interpreting the composition of the dataset.

Reference-data quality is also a research-design issue rather than only an operational concern. Bottega and Powell [25] argue for a universal legal entity identifier to improve financial data linkage, and the Financial Stability Board [26] sets out the policy rationale for a global LEI framework. At the instrument level, ISO 6166:2013 [27] standardizes the ISIN system, while ISO 10962:2015 [28] standardizes the CFI classification used to describe security characteristics across markets.

Complementary identifier and classification frameworks further improve interoperability. Bloomberg's FIGI white paper [29] presents an open symbology approach for mapping instruments across systems, and the MSCI and S&P Dow Jones Indices GICS methodology [30] provides a structured sector and industry taxonomy. This literature explains why strong ISIN, SEDOL, CFI, and FIGI coverage supports security-level matching, whereas weaker LEI and GICS coverage constrains issuer-level and sector-based inference unless the file is supplemented externally.

Taken together, the literature from 2010 to 2020 supports treating the source file as a pre-2020 securities reference layer for comparative research on market composition, venue concentration, ETF structure, and identifier completeness. The market-quality literature explains why venue fields matter, the cross-listing literature explains why country and currency fields matter, the ETF literature explains why instrument-type separation matters, and the standards literature clarifies why identifier completeness is central to reproducibility.

The present study therefore positions the dataset not as a passive administrative register, but as research infrastructure whose empirical value depends on how comprehensively it captures instruments, venues, and linkable identifiers. That framing allows the analysis to connect descriptive evidence to broader questions of dataset usability in international finance research.

Table 1 presents all studies cited in the literature review and shows how each individual source contributes to the analytical framing of the present study.

Table 1. Study-by-study literature review.

Ref.	Study	Main focus	Relevance to the present study
[1]	Indriawan (2020)	Macroeconomic news and market quality in the Australian stock market.	Shows that market quality varies with information arrival, supporting attention to trading environments and venue conditions.
[2]	Brolley (2020)	Execution risk and price improvement across lit and dark markets.	Supports interpreting venue structure as economically meaningful rather than purely administrative metadata.
[3]	Hendershott, Jones, and Menkveld (2011)	Effect of algorithmic trading on quoted liquidity.	Links trading technology to liquidity quality, reinforcing why exchange context matters in reference-data analysis.
[4]	Menkveld (2013)	Role of high-frequency traders in modern market making.	Motivates the treatment of venue design as part of market structure.
[5]	Brogaard, Hendershott, and Riordan (2014)	High-frequency trading and price discovery.	Shows that trading venues influence information incorporation and market quality.
[6]	Budish, Cramton, and Shim (2015)	Speed competition and market-design distortions under continuous limit order books.	Supports the broader literature stream connecting exchange design to trading outcomes.

Ref.	Study	Main focus	Relevance to the present study
[7]	O'Hara (2015)	Synthesis of high-frequency market microstructure.	Provides conceptual grounding for using venue and trading fields in comparative analysis.
[8]	Foley and Putnins (2016)	Dark trading and market quality under different conditions.	Reinforces that market outcomes depend on venue structure and transparency conditions.
[9]	Athanassakos et al. (2010)	Determinants of investor demand for cross-listed firms.	Supports the importance of listing location in shaping market participation and visibility.
[10]	Doidge, Karolyi, and Stulz (2013)	Shift of equity issuance activity outside the United States.	Helps frame listing-country evidence within global capital-market geography.
[11]	Dodd (2013)	Review of cross-listing theories.	Provides theory for interpreting country and exchange choices in an international sample.
[12]	Li, Brockman, and Zurbruegg (2015)	Cross-listing, information environment, and corporate governance.	Supports using country and venue variables to interpret cross-market differences.
[13]	Sarkissian and Schill (2016)	Cross-listing waves over time.	Shows that listing activity is patterned rather than random, strengthening interpretation of country concentration.
[14]	Cumming, Hou, and Wu (2018)	Exchange rules, governance, and trading location of cross-listed stocks.	Directly links exchange governance to venue choice, supporting country and exchange analysis.
[15]	IOSCO (2013)	Regulatory principles for exchange-traded funds.	Provides institutional background for understanding ETF structure and market role.
[16]	U.S. Securities and Exchange Commission (2019)	Regulatory description of exchange-traded funds.	Clarifies ETF operating mechanisms relevant to interpreting the ETF-heavy sample.
[17]	Madhavan and Sobczyk (2016)	ETF price dynamics and liquidity.	Supports separating ETFs from other securities when interpreting market composition.
[18]	Israeli, Lee, and Sridharan (2017)	Information effects associated with ETF trading and ownership.	Explains why ETF prevalence may affect market behavior and interpretation.
[19]	Petajisto (2017)	Pricing inefficiencies in exchange-traded funds.	Shows that ETFs have distinctive pricing dynamics that justify explicit treatment in the study.
[20]	Da and Shive (2018)	ETF trading and asset-return comovement.	Reinforces the analytical importance of ETF presence in a multi-market sample.

Ref.	Study	Main focus	Relevance to the present study
[21]	Ben-David, Franzoni, and Moussawi (2018)	Volatility effects associated with ETFs.	Supports interpreting ETF concentration as a substantive market-structure feature.
[22]	Broman and Shum (2018)	ETF flows, relative liquidity, and short-term demand.	Further motivates distinguishing ETF instruments from operating-company equity.
[23]	Box, Davis, and Fuller (2019)	ETF competition and market quality.	Connects ETF growth to broader market-quality questions relevant to comparative reference data.
[24]	Laipply and Madhavan (2020)	Pricing and liquidity of fixed-income ETFs during the 2020 crisis.	Shows ETF structure matters for resilience and liquidity interpretation.
[25]	Bottega and Powell (2011)	Universal legal entity identifier as a linchpin for financial data.	Justifies the manuscript's emphasis on identifier completeness as a research-design issue.
[26]	Financial Stability Board (2012)	Policy rationale for a global LEI framework.	Supports evaluating entity-linkage quality rather than assuming identifier sufficiency.
[27]	ISO 6166:2013	Standardization of the ISIN system.	Explains why ISIN completeness is central to reproducible cross-market research.
[28]	ISO 10962:2015	Standardization of the CFI classification system.	Supports the assessment of classification-field completeness and usability.
[29]	Bloomberg / OpenFIGI white paper (2015)	Open symbology and financial instrument mapping.	Motivates evaluating FIGI-related fields for interoperability and linkage.
[30]	MSCI and S&P Dow Jones Indices GICS methodology (2018)	Sector and industry classification framework.	Clarifies why weak GICS coverage limits sector-based inference in the dataset.

III. RESEARCH QUESTIONS

- What is the composition of the pre-2020 sample by security type, listing country, exchange, and trading currency?
- How concentrated is the dataset across major venues and currencies, and what does that imply for its value as a global pre-2020 reference file?
- To what extent are the core identifier and classification fields complete enough to support reproducible cross-market research?

IV. METHODOLOGY

1. DATA SOURCE AND SAMPLE CONSTRUCTION

The study uses a global listed-securities reference file as its sole data source. The source table contains 237 records and 56 fields organized across three worksheet blocks. To align the analysis with the stated temporal

scope, all records created in 2020 or later were excluded, leaving a retained analytical sample of 147 pre-2020 observations.

The dataset includes security descriptions, issuer names, ISIN, FIGI, SEDOL, CFI, LEI, GICS, exchange codes, listing-country codes, trading currencies, lot sizes, and shares outstanding. This structure allows the study to assess both the composition of listed instruments and the completeness of the identifiers required for replication, linkage, and classification.

2. VARIABLES AND MEASURES

Variables are grouped into four categories. The first captures instrument and issuer identification, including ISIN, FIGI, SEDOL, issuer name, and security description. The second covers classification fields such as CFI, CIC, GICS, and security type. The third includes market-location variables, notably primary exchange, listing country, and trading currency. The fourth consists of quantitative attributes such as lot size and shares outstanding.

These variables make it possible to evaluate both substantive market composition and the practical usability of the dataset for downstream empirical analysis.

3. ANALYTICAL STRATEGY

The empirical strategy combines descriptive frequency analysis, cross-tabulation, date filtering, and key-field completeness checks. Country, exchange, security-type, and currency distributions are reported as counts and percentages. Temporal coverage is summarized by record-creation year. Lot size and shares outstanding are described with robust statistics, including quartiles and medians, because both variables are highly right-skewed.

Field completeness is evaluated for the identifiers and classifications most relevant to future database linkage and market-structure analysis. Because the emphasis is on dataset characterization rather than hypothesis testing, the design is descriptive rather than causal.

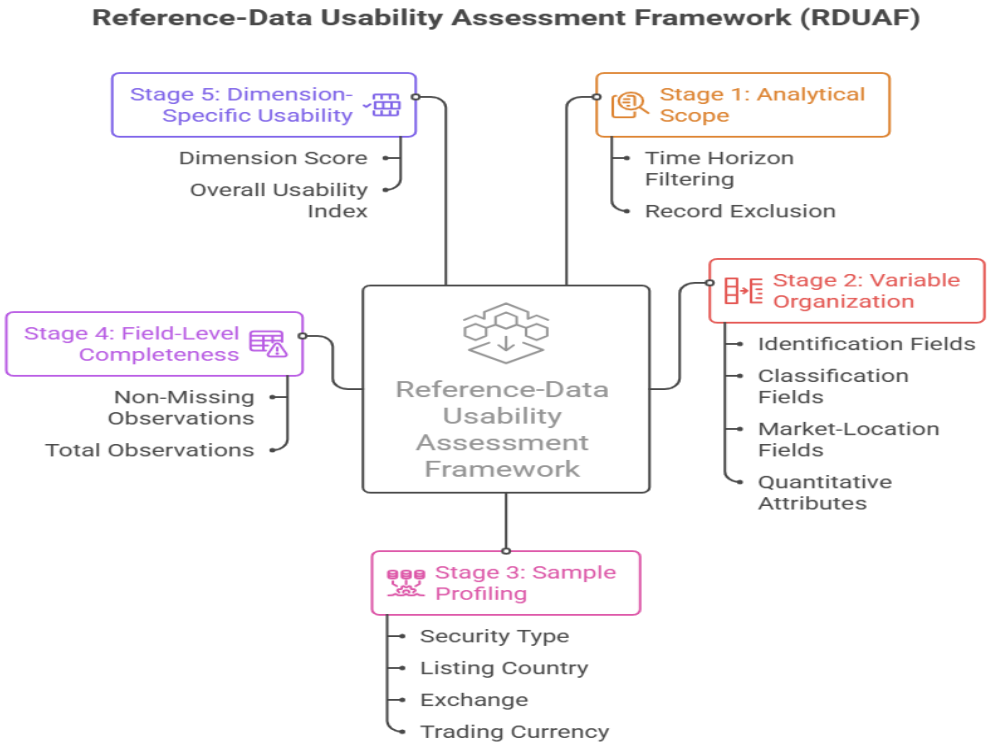
4. PROPOSED METHOD: REFERENCE-DATA USABILITY ASSESSMENT FRAMEWORK

Component	Description
Purpose	The present study proposes a Reference-Data Usability Assessment Framework (RDUAF) for assessing whether a securities reference dataset is suitable for a specific empirical design. The method is intended as a pre-analysis screening framework rather than as a predictive model. Its purpose is to transform descriptive evidence on data structure and completeness into a transparent research-usability judgment.
Stage 1	Defines the analytical scope by filtering the raw file to the target time horizon and excluding records that fall outside the study period.
Stage 2	Maps the retained variables into four dimensions: identification fields, classification fields, market-location fields, and quantitative attributes.
Stage 3	Profiles the retained sample using counts, percentages, and concentration patterns across security type, listing country, exchange, and trading currency.
Stage 4	Evaluates field-level completeness. For each field j, completeness is defined as $C_j = (n_j / N) \times 100$, where n_j is the number of non-missing observations and N is the total number of retained observations.
Stage 5	Converts field-level results into dimension-specific usability judgments. For a dimension k containing m_k fields, the dimension score is defined as $D_k = (1 / m_k) \Sigma(C_j)$, and the overall usability index may be expressed as $OUI = \Sigma(w_k D_k)$, where w_k represents researcher-defined weights reflecting the intended use of the dataset.
Output and interpretation	In the present article, the proposed method is used as a structured interpretive framework rather than as a standalone classification algorithm. Its practical output is a

	three-part judgment covering suitability for security-level descriptive research, suitability for issuer-level linkage, and suitability for sector-based analysis. This design is consistent with the empirical evidence reported later in the paper, where core security identifiers are strong but LEI and GICS completeness remain materially weaker.
Relation to Figure 1	Figure 1 summarizes the main stages of the study, from sample construction through interpretation of dataset usability, and therefore illustrates the operational sequence of the proposed RDUAF method.

FIGURE 1. Study workflow and analytical stages.

The workflow highlights that the design is descriptive and sequential: the retained pre-2020 sample is first defined, variables are organized, descriptive evidence is generated, and data-quality implications are interpreted. In that sense, the workflow serves both as a summary of the empirical procedure and as a visual representation of the proposed reference-data usability method.



5. SCOPE, VALIDITY, AND REPRODUCIBILITY

The analysis is intentionally limited to the information contained in the source file. The study does not use transaction-level, price-series, or return data, and it therefore does not attempt causal claims about liquidity, efficiency, or valuation. Instead, it evaluates whether the structure and completeness of the file support rigorous descriptive and comparative research.

Reproducibility depends on transparent reporting of sample filters, variable definitions, and missing-data patterns. Missingness is interpreted substantively because incomplete identifiers can constrain sample construction and weaken downstream inference. The dataset contains no personal data, so no human-subject ethics issues arise in the present analysis.

V. RESULTS

The retained pre-2020 sample contains 147 of the 237 raw records after the year filter is applied. Across the sample, ETFs and common equity dominate the instrument mix, listings are concentrated in a limited set of countries and venues, and identifier coverage is strongest for security-level fields while materially weaker for issuer-level and sector-level variables.

1. SAMPLE CONSTRUCTION AND SCOPE

Table 2. Sample construction and scope.

Metric	Value
Raw records in source file	237
Records retained for pre-2020 analysis	147
Records excluded (2020 and later)	90
Fields in source table	56
Worksheet blocks represented	3
Listing countries represented	11

Table 2 establishes the analytical scope of the study. After excluding records created in 2020 or later, 147 observations remain for analysis. The retained sample spans 11 listing countries and preserves the full 56-field structure of the source table, providing sufficient breadth for descriptive cross-market comparison.

2. GEOGRAPHIC COVERAGE AND SOURCE-FILE STRUCTURE

Table 3. Composition by source worksheet block.

Worksheet block	Count	Share (%)
Asia worksheet block	87	59.2
Europe worksheet block	52	35.4
Africa worksheet block	8	5.4

The worksheet blocks describe how the source file is organized rather than a strict geographic taxonomy. The block labeled Asia is the largest, but because it contains listings beyond a single region, country-level evidence provides the more reliable basis for substantive interpretation.

Table 4. Listing country distribution.

Listing country	Count	Share (%)
GB	24	16.3
AU	21	14.3
HK	20	13.6
US	18	12.2
NL	16	10.9
CA	14	9.5
IT	12	8.2
MX	9	6.1
ZA	6	4.1
BR	5	3.4
MA	2	1.4

The sample is international in coverage but not evenly distributed across markets. GB, AU, HK, US, and NL together account for 99 of the 147 retained observations, or 67.3% of the sample. The evidence therefore points to meaningful cross-market breadth alongside clear geographic concentration.

3. INSTRUMENT COMPOSITION

Table 5. Security type mix.

Security type	Count	Share (%)
ETF	79	53.7
EQS	64	43.5
PRF	2	1.4
DR	1	0.7
DST	1	0.7

ETFs represent the largest share of the retained sample, followed by common equity. Together these two categories account for 97.2% of observations, indicating that the file functions primarily as a pre-2020 ETF-and-equity reference dataset. Preferred shares, depository receipts, and trusts appear only marginally and should not drive the interpretation of the sample as a whole.

Figure 2 visually reinforces the dominance of ETFs and common equity in the retained sample.

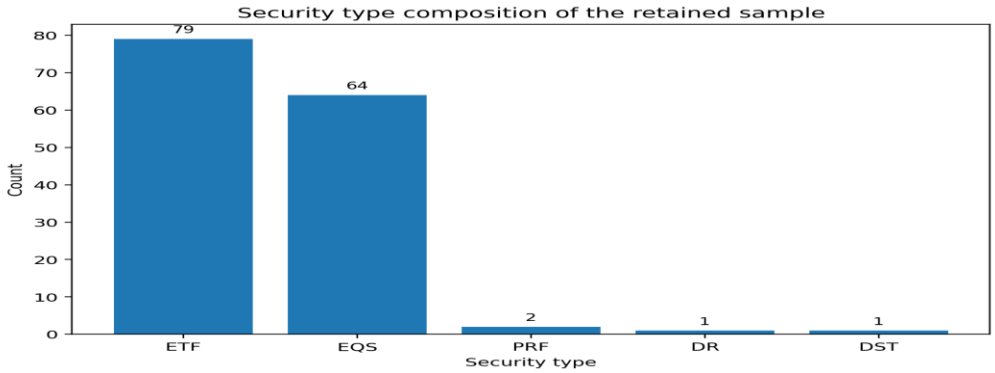


FIGURE 2. Security type composition of the retained sample.

The figure complements Table 5 by showing that ETFs and common equity account for nearly the entire sample, with all other security types appearing only marginally.

4. TRADING VENUES AND CURRENCIES

Table 6. Top primary exchanges.

Primary exchange	Count	Share (%)
USNASD	38	25.9
AUASX	21	14.3
HKSEHK	20	13.6
CACNQ	10	6.8
USNYSE	10	6.8
CHSSX	10	6.8
ITMSE	8	5.4
ZAJSE	6	4.1
DEFSX	6	4.1
BRBVSP	5	3.4

NASDAQ is the single largest venue code in the retained sample, followed by the Australian and Hong Kong exchanges. The top three venue codes account for 79 records, or 53.7% of the sample, which indicates meaningful concentration despite the cross-country span of the dataset.

Table 7. Trading currency distribution.

Trading currency	Count	Share (%)
USD	42	28.6
EUR	29	19.7
AUD	21	14.3
HKD	17	11.6
CAD	14	9.5
MXN	8	5.4
ZAR	6	4.1
BRL	5	3.4
CNY	2	1.4
MAD	2	1.4
GBP	1	0.7

Currency exposure is similarly concentrated. USD and EUR together account for 48.3% of retained observations, while the four largest currencies—USD, EUR, AUD, and HKD—account for 74.1%. The file is therefore internationally diversified but still centered on a small number of dominant trading currencies.

5. TEMPORAL COVERAGE WITHIN THE RETAINED SAMPLE

Table 8. Record-creation year counts.

Year	Records added
2001	17
2002	4
2003	2
2004	2
2005	1
2006	5
2007	6
2008	3
2010	3
2011	4
2012	6
2013	7
2014	4
2015	16
2016	4
2017	14
2018	37
2019	12

The retained observations span 2001 to 2019. Record additions intensify in the later years of the sample, particularly in 2015, 2017, 2018, and 2019, with 2018 alone contributing 37 records. The dataset therefore has a clear pre-2020 temporal profile while also showing stronger accumulation in the second half of the covered period.

Table 9. Robust summary of key numeric fields.

Field	Non-missing n	Min	25th pct	Median	75th pct	Max
Lot size	110	1	1	1	100	20,000
Shares outst.	146	27,350	15,886,792	50,600,267	186,369,440	24,780,937,000

Median values are more informative than means because both numeric fields are strongly right-skewed. Lot size is typically minimal, whereas shares outstanding vary substantially across instruments and issuers. The broad range of shares outstanding suggests that the sample combines products and firms of very different scale, reinforcing the need for careful segmentation in subsequent empirical work.

6. IDENTIFIER COMPLETENESS AND DATA QUALITY

Table 10. Completeness of key fields.

Field	Non-missing	Completeness (%)
ISIN	147	100.0
IssuerName	147	100.0
SecurityDescription	147	100.0
SEDOL	147	100.0
TradingCurenCD	147	100.0
CFI	146	99.3
CIC	146	99.3
PrimaryExchgCD	146	99.3
FigiGlobalShareclassID	145	98.6
FIGI	141	95.9
LEI	68	46.3
GICS	60	40.8

Security-level identifiers are the strongest part of the dataset. ISIN, SEDOL, issuer name, security description, and trading currency are complete for all retained records, while CFI, CIC, primary exchange, and the FIGI share-class field are also nearly complete. By contrast, LEI and GICS coverage is substantially weaker, which limits issuer-level linkage and sector-based inference unless the file is enriched with external sources.

Figure 3 provides a visual summary of field completeness and highlights the contrast between strong security-level identifiers and weaker issuer- and sector-level fields.

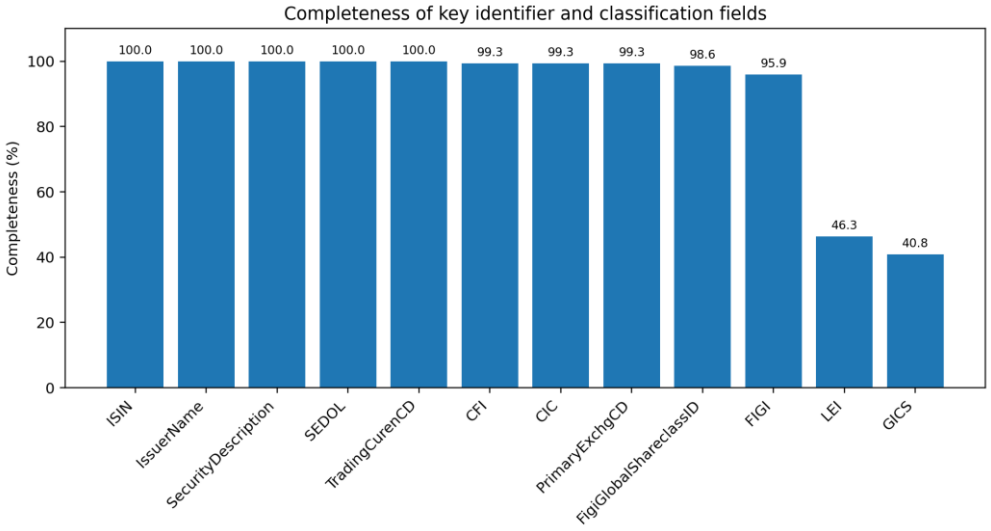


FIGURE 3. Completeness of key identifier and classification fields.

The visual pattern makes clear that most security-level fields are effectively complete, whereas LEI and GICS are materially less complete and therefore constrain issuer-level and sector-based analysis.

VI. DISCUSSION

1. WHAT THE DATASET CAPTURES RELIABLY

The results suggest that the study is best understood as a pre-2020 security master dataset suited to descriptive and comparative analysis rather than to causal estimation. Its strongest analytical value lies in instrument identification and market profiling. The dominance of ETFs and common equity indicates that the file captures the core exchange-traded product universe represented in the source table, while the marginal presence of other instrument classes shows that those categories remain peripheral.

This product mix matters for interpretation. Pooled analyses that ignore the ETF-heavy composition of the sample may blur important differences in structure, liquidity, and economic meaning between funds and operating firms. The dataset is therefore more useful when instrument types are separated explicitly rather than treated as interchangeable observations.

2. IMPLICATIONS FOR CROSS-MARKET RESEARCH DESIGN

The concentration of records in a limited set of countries, exchanges, and currencies is substantively meaningful. GB, AU, HK, US, and NL dominate the country profile, while the top exchanges and currencies account for a large share of observations. This pattern is consistent with the literature on international listing choice and venue concentration, which emphasizes market depth, investor access, and institutional visibility in shaping where securities are listed and traded.

3. DATA-QUALITY IMPLICATIONS AND STUDY LIMITATIONS

The data-quality results are equally consequential. Strong coverage for ISIN, SEDOL, FIGI-related fields, issuer names, and security descriptions makes the dataset a credible base for security-level matching and descriptive profiling across venues. However, the much weaker coverage of LEI and GICS constrains issuer-level linkage and sector-based analysis. Studies that depend on firm-level entity resolution, ownership integration, or sector controls would require supplementary enrichment before strong conclusions could be drawn.

Several limitations qualify the findings. The sample is relatively small, it reflects the structure of one source master file rather than the entire global listed universe, and its worksheet blocks are organizational labels rather than a formal regional taxonomy. In addition, because the study relies on reference-data fields rather than returns, quotes, or order-flow measures, it cannot directly test market efficiency, transaction costs, or causal price effects. Its contribution is instead to clarify what this dataset can support reliably as a basis for subsequent empirical research.

VII. CONCLUSION

This study reframes a pre-2020 global listed-securities file as research infrastructure rather than as a purely administrative register. The evidence shows a dataset dominated by ETFs and common equity, concentrated in a limited set of countries, venues, and currencies, and supported by very strong coverage for core security identifiers.

At the same time, materially weaker LEI and GICS coverage constrains issuer-level and sector-based analysis unless supplementary sources are added. The paper therefore contributes a clearer analytical framing, a transparent methodological structure, and a practical assessment of where securities reference data are most useful in empirical finance.

Beyond its empirical findings, the study offers RDUAF as a reusable methodological template for researchers who need to evaluate whether securities reference data are adequate for descriptive profiling, market-composition analysis, identifier-based linkage, or sector-based designs before more advanced empirical modeling is attempted.

Future research can extend this dataset by linking it to returns, trading records, ownership information, or firm-level fundamentals in order to examine how reference-data quality influences broader empirical inference across markets.

Author Contributions

The author conducted the conceptualization, methodology, data analysis, investigation, writing, review, editing, and final approval of the manuscript.

Funding

This research received no external funding.

Data Availability

The dataset will be available from the author upon reasonable request.

Conflicts of Interest

The author declares no conflict of interest.

REFERENCES

1. Indriawan, I. (2020). Market quality around macroeconomic news announcements: Evidence from the Australian stock market. Pacific-Basin Finance Journal, 61, 101071.
2. Brolley, M. (2020). Price improvement and execution risk in lit and dark markets. Management Science, 66(2), 863-886.
3. Hendershott, T., Jones, C. M., & Menkveld, A. J. (2011). Does algorithmic trading improve liquidity? The Journal of Finance, 66(1), 1-33.
4. Menkveld, A. J. (2013). High frequency trading and the new-market makers. Journal of Financial Markets, 16(4), 712-740.
5. Brogaard, J., Hendershott, T., & Riordan, R. (2014). High-frequency trading and price discovery. Review of Financial Studies, 27(8), 2267-2306.
6. Budish, E., Cramton, P., & Shim, J. (2015). The high-frequency trading arms race: Frequent batch auctions as a market design response. Quarterly Journal of Economics, 130(4), 1547-1621.
7. O'Hara, M. (2015). High frequency market microstructure. Journal of Financial Economics, 116(2), 257-270.

8. Foley, S., & Putnins, T. J. (2016). Should we be afraid of the dark? Dark trading and market quality. Journal of Financial Economics, 122(3), 456-481.
9. Athanassakos, G., Ackert, L. F., Naydenova, B., & Tafkov, I. D. (2010). Determinants of investor demand for cross-listed firms. Financial Markets, Institutions & Instruments, 19(3), 245-267.
10. Doidge, C., Karolyi, G. A., & Stulz, R. M. (2013). The U.S. left behind? Financial globalization and the rise of IPOs outside the U.S. Journal of Financial Economics, 110(3), 546-573.
11. Dodd, O. (2013). Why do firms cross-list their shares on foreign exchanges? A review of cross-listing theories and empirical evidence. Review of Behavioral Finance, 5(1), 77-99.
12. Li, S., Brockman, P., & Zurbruegg, R. (2015). Cross-listing, firm-specific information, and corporate governance: Evidence from Chinese A-shares and H-shares. Journal of Corporate Finance, 32, 347-362.
13. Sarkissian, S., & Schill, M. J. (2016). Cross-listing waves. Journal of Financial and Quantitative Analysis, 51(1), 259-306.
14. Cumming, D., Hou, W., & Wu, E. (2018). Exchange trading rules, governance, and trading location of cross-listed stocks. The European Journal of Finance, 24(16), 1453-1484.
15. International Organization of Securities Commissions. (2013). Principles for the regulation of exchange traded funds. Final Report. IOSCO.
16. U.S. Securities and Exchange Commission. (2019). Exchange-Traded Funds. Investment Company Act Release No. IC-33646.
17. Madhavan, A., & Sobczyk, A. (2016). Price dynamics and liquidity of exchange-traded funds. Journal of Investment Management, 14(2), 1-17.
18. Israeli, D., Lee, C. M. C., & Sridharan, S. A. (2017). Is there a dark side to exchange traded funds? An information perspective. Review of Accounting Studies, 22(3), 1048-1083.
19. Petajisto, A. (2017). Inefficiencies in the pricing of exchange-traded funds. Financial Analysts Journal, 73(1), 24-54.
20. Da, Z., & Shive, S. (2018). Exchange traded funds and asset return correlations. European Financial Management, 24(1), 136-168.
21. Ben-David, I., Franzoni, F., & Moussawi, R. (2018). Do ETFs increase volatility? The Journal of Finance, 73(6), 2471-2535.
22. Broman, M. S., & Shum, P. M. (2018). Relative liquidity, fund flows and short-term demand: Evidence from exchange-traded funds. The Financial Review, 53(1), 87-115.
23. Box, T., Davis, R. L., & Fuller, K. P. (2019). ETF competition and market quality. Financial Management, 48(3), 873-916.
24. Laipply, S., & Madhavan, A. (2020). Pricing and liquidity of fixed income ETFs in the Covid-19 virus crisis of 2020. The Journal of Index Investing, 11(3), 7-19.
25. Bottega, J. A., & Powell, L. F. (2011). Creating a linchpin for financial data: Toward a universal legal entity identifier. Journal of Economics and Business, 63(6), 105-115.
26. Financial Stability Board. (2012). A global legal entity identifier for financial markets. FSB.
27. International Organization for Standardization. (2013). ISO 6166:2013 Securities and related financial instruments - International securities identification numbering system (ISIN). ISO.
28. International Organization for Standardization. (2015). ISO 10962:2015 Securities and related financial instruments - Classification of financial instruments (CFI code). ISO.
29. Bloomberg L.P. (2015). The case for open symbology: Financial Instrument Global Identifier (FIGI) white paper. Bloomberg / OpenFIGI.
30. MSCI & S&P Dow Jones Indices. (2018). Global Industry Classification Standard (GICS) methodology. MSCI and S&P Dow Jones Indices.