


Luxembourg Fund Data Repository

Angeliki Skoura * , Julian Presber and Jang Schiltz

Department of Finance, University of Luxembourg, 4365 Esch-sur-Alzette, Luxembourg; presber@pt.lu (J.P.); jang.schiltz@uni.lu (J.S.)

* Correspondence: angeliki.skoura@uni.lu; Tel.: +352-466-644-5475

Received: 16 June 2020; Accepted: 16 July 2020; Published: 19 July 2020



Abstract: In this paper, we introduce the Luxembourg Fund Data Repository, a novel database of investment funds available for academic research that was created at the Department of Finance of the University of Luxembourg. The database contains the population of Undertakings for Collective Investment in Transferable Securities funds domiciled in Luxembourg from the starting month of their existence (March 1988) to October 2016. The fund characteristics are organized in a comprehensive database architecture encompassing static and dynamic data over the entire life of the funds. The characteristics include fund identifiers, official name, status information, management company and other service providers, daily and monthly performance time-series, portfolio holdings, classification of investment objective, fees, dividends, and cash flows. The database was constructed after collecting and assembling complementary historical information from three data providers. Importantly, funds no longer in existence due to liquidation or mergers are included in the database, preventing survivorship bias. The database has been constructed to serve as a research dataset of high accuracy due to the maximization of population coverage, the maximization of historical coverage, and validation by using information acquired from the supervisory authority of the financial sector of Luxembourg. License currently available to researchers of the Department of Finance of the University of Luxembourg. Future plans for extending accessibility to the global academic community.

Keywords: investment fund data; fund time-series; Undertakings for Collective Investment in Transferable Securities; survivorship-bias-free database

1. Summary

Investors across the globe have demonstrated strong demand for regulated open-end funds in the past decade, and total net assets of worldwide regulated open-end funds totaled \$46.7 trillion at the end of 2018 [1]. The total net assets under the management of European investment funds reached €15.2 trillion in 2018 according to the European Fund and Asset Management Association [2]. Though Luxembourg is the leading investment fund domicile in Europe and the second largest worldwide behind the United States (US), with €4404 billion assets under management as of March 2019 [3], academic researchers in the investment fund literature have so far been working primarily with US mutual fund data distributed by the Center for Research in Security Prices (CRSP) [4,5]. The investment fund industry of Luxembourg is a worldwide leader in cross-border fund distribution. Luxembourg-domiciled investment funds are distributed in more than 75 countries around the globe, with a particular focus on Europe, Asia, Latin America, and the Middle East [6]. The largest segment of investment funds in Luxembourg is comprised of the Undertakings for Collective Investment in Transferable Securities (UCITS). UCITS are investment vehicles that invest in liquid assets and can be publicly marketed and sold to retail investors throughout the European Union (EU). Today, UCITS funds are the most widely accepted retail investment funds worldwide and constitute a well-regulated investment product with significant levels of investor protection [3]. According to the

authors of [7], UCITS are currently recognized as a gold standard of investor's protection. This category of investment funds accounts for around 75% of all collective investments by small investors in Europe, as of 2019 [8]. Unlike US mutual funds, UCITS have been designed to be marketed cross-border because it is currently possible to sell shares of UCITS not only within but also beyond European borders [7]. They are currently available to the vast majority of non-US retail and institutional investors. The concept of a harmonized European investment fund product was originally introduced in the European Directive 85/611/EEC of 20 December 1985, which provided a single regulatory regime across the EU for open-ended funds investing in transferable securities such as shares and bonds. With a view to achieving the highest levels of investor protection, further European directives on UCITS were adopted to regulate the organization, management, and oversight of such funds, and they impose rules concerning diversification, liquidity, and they use of leverage. Funds that meet the conditions laid down by the above-mentioned directives are commonly known as UCITS. Luxembourg was the first European Union Member State to pass UCITS legislation into national law in 1988; today, UCITS represent the majority of investment funds in the country's fund industry. Luxembourg-domiciled UCITS constitute more than 65% of the internationally distributed UCITS [9]. As of December 2019, Luxembourg is the leading pan-European distribution platform for UCITS, and the fund assets under management of the UCITS market in Luxembourg is €3920.8 billion, which represents 35.7% of the total UCITS market [10]. Considering the size and diversity of the investment fund industry of Luxembourg, the country offers a unique case for the development of a new database in terms of academic research on investment funds.

This paper presents the Luxembourg Fund Data Repository (LFDR), a new research database¹ that has recently been developed at the Department of Finance of the University of Luxembourg. The LFDR contains UCITS funds domiciled in Luxembourg from March 1988, which was the first month of UCITS' existence, until October 2016. The LFDR was constructed to serve as a research dataset of high quality due to three key features: its validation using data from official data sources of Luxembourg, the maximization of population coverage (the inclusion of both active and obsolete UCITS and the provision of appropriate referencing for merged funds), and the maximization of historical coverage. The included data fields were organized in a comprehensive database structure that reflects the building blocks of UCITS in terms of their governing European and national UCITS legislation. The structural characteristics of UCITS can be found in Appendix A. The data fields represent static and dynamic characteristics that lend themselves to both cross-sectional analysis across funds and longitudinal analysis in a time-series. The field categories include, *inter alia*, identifiers, official name, status information such as active or obsolete, classification of investment objective, fees, time-series of net asset value, total net assets, return, dividends, portfolio holdings, and cash flows. Additional fields describing the structural particularities of UCITS funds such as the internal parent-child relationships inside an umbrella fund are also represented in the database. The data were populated after assembling complementary historical information collected from two commercial data providers, namely Morningstar and Fundsquare², and were validated and enhanced by using data provided by the supervisory authority of the Luxembourg financial sector (Commission de Surveillance du Secteur Financier—CSSF). The current version of the database is open to the Department of Finance researchers at the University of Luxembourg under license restriction and only for academic research purposes. The full license agreement that researchers need to sign in order to access the database is presented in Appendix B. The future development plans of the database include regular updates of the current content with more recent data and the extension of the database accessibility to the global academic community.

The data of the LFDR permit an analysis of factors related to underlying portfolio construction, investment flows, and portfolio performance, including specific aspects of performance such as fees.

¹ For the rest of the paper, the terms "the LFDR" and "the database" are used interchangeably.

² Fundsquare is a subsidiary of Luxembourg Stock Exchange.

All these factors can together be seen as a proxy for the study of investment funds as a product of investment activity. These factors are similar to those offered in the database of its counterpart in the United States, the CRSP Mutual Fund database, but they represent an important innovation with respect to UCITS databases. Additionally, the LFDR contains a more globally representative set of funds as a proxy for international investment flows than the CRSP Mutual Fund database. The latter contains a strong US-centric bias, as the vast majority of both underlying portfolio investments and subscribing investors are located in US. In contrast, the UCITS represent a set of investment funds that are globally diversified in terms of both their portfolio investments and their underlying unit-holder base (recognizing, however, that there are usually no US-based investors in UCITS). This greater level of representativity of global investment activity constitutes an important new addition to the study of investment funds and international investment, and it is a key aspect of the innovation that the LFDR brings.

2. Data Description

The database contains data for Luxembourg-domiciled UCITS covering the period from March 1988 to October 2016. The database contains both active and obsolete (liquidated or merged) UCITS, including appropriate referencing for merged funds. The database architecture reflects the structure of UCITS funds. Details about the structure and the main characteristics of UCITS with respect to fund, sub-fund(s), and share class(es) can be found in Appendix A. Section 2.1 describes the number of entities in the database in terms of funds, sub-funds, and share classes. Section 2.2 presents the data tables and the data fields that were created to organize the data of funds, sub-funds, and share classes. Section 2.3 focuses on the historical coverage of the major time-series fields by presenting figures of the distribution of historical observations per year. Finally, Section 2.4 provides details on how the core data tables are linked and on how the parent–child relationships of UCITS parts are represented in the database.

2.1. Database Population

Table 1 presents the database population in terms of total number of funds, sub-funds, and share classes. The database contains 4591 unique funds, 18,982 unique sub-funds, and 84,556 unique share classes. The status information such as active or obsolete status of the above-mentioned entities at the end of October 2016 is also presented in the table. In addition, there are 14,464 unique portfolios associated with the sub-funds. Figure 1 illustrates the break-down of the total fund population per year of fund's constitution date since 1988. It is worth mentioning that in the first year of UCITS existence (1988), 65 new UCITS were created and 180 fund entities, which had been constituted before, were transformed into UCITS in that year.

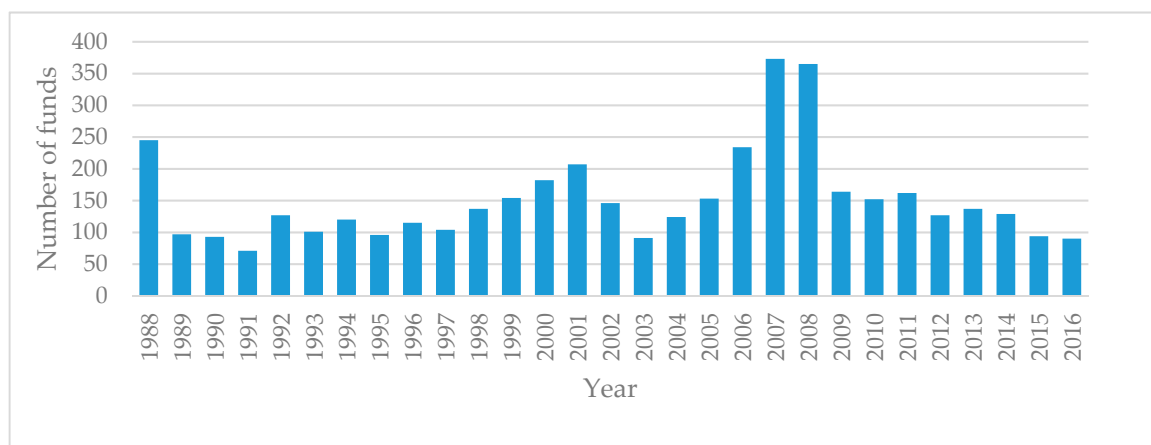


Figure 1. The number of fund entities grouped by year of constitution date.

Table 1. The total number of funds, sub-funds, and share classes in the database, along with their status (either active or obsolete as of 31 October 2016).

	Total Number	Active	Obsolete
<i>Fund</i>	4591	2677	1914
<i>Sub-Funds</i>	18,982	10,017	8965
<i>Share Classes</i>	84,556	54,566	29,990

2.2. Data Tables and Data Fields

The content of the database is organized in data tables. Each data table groups together data fields that are conceptually related. In total, there are 24 data tables that include data fields related to four information groups: four data tables related to the fund level, seven data tables related to the sub-fund level, 10 data tables related to the share class level, and three data tables related to portfolio data. Before presenting the underlying data fields of the data tables, the identifiers of the core entities are reported here; each single fund, sub-fund, share class entity, and portfolio entity is associated with a unique proprietary identifier (FundID, SubfundID, ShareclassID, and PortfolioID, respectively). The data tables related to the fund level include fields such as fund identifiers (including FundID), official fund name, status (active or obsolete), constitution date, legal form, management company, auditor company, and custodian and contact information of the fund management company. The data tables related to the sub-fund level include fields such as sub-fund identifiers (including SubfundID), official sub-fund name, status (active or obsolete), launch date, daily time-series of total net assets (TNA), classification of investment objective based on prospectus, classification of investment objective based on portfolio assets, and monthly time-series of cash flows (namely the amount of net subscriptions and net redemptions). The data tables related to the share class level include fields such as share class identifiers including (including ShareclassID), official share class name, status (active or obsolete), launch date, subscription fee, redemption fee, daily time-series of TNA, daily time-series of net asset value (NAV), monthly time-series of return, time-series of dividends, and classification of investment objective in case of currency-hedged share classes. Finally, the data tables related to portfolio information include fields such as portfolio identifiers (including PortfolioID), list of portfolio report dates, complete list of portfolio holdings and asset allocation in terms of predefined equity sub-categories, predefined bond sub-categories, and cash.

Associating specific values of fields with effective date(s) is required in the database. Some fields change their values on a regular basis; these include NAV, which is updated daily, while other fields may change on a non-regular basis such as the fund name. Furthermore, some fields may change their value a maximum of once during the fund life, such as the end date of a fund. In order to organize the time information needed for data fields in a comprehensive and storage efficient way, the data tables are classified in the following three categories:

1. Data tables of life-cycle fields: These data tables contain fields that are life-cycle constant or may change a maximum of once during the entire life of the fund. Thus, one date maximum is associated with the values of these fields.
2. Data tables of period-based fields: These data tables contain fields that change on a non-regular basis. Thus, each entry of these data tables contains the start date and the end date of the effective period.
3. Data tables of time-series fields: These data tables contain fields that change on a periodic basis. The fields constitute time-series with a constant frequency. Thus, each entry of these data tables contains one effective date.

Table 2 provides an overview of the database content, including the above-mentioned classification of data tables. The first column (from left to right) of Table 2 presents the four main information groups related to UCITS: fund, sub-fund, portfolio, and share class. The second column presents the

names of the 24 data tables. The third column presents the complete list of data fields per data table, while the fourth column presents the category of each data table as described in the paragraph above. The last column reports the number of entries of each data table; furthermore, in case of period-based and time-series fields, the history range of data availability is denoted in parentheses. For example, let us consider the data table with the name “Subfund.DailySeries.” Each entry of this table contains the following fields: SubfundID (unique identifier of sub-fund entity), TNA (amount of total net assets for the sub-fund), date (the effective date of TNA), and currency (the currency in which TNA is expressed). The complete definitions of the fields are documented in the user manual that accompanies the database. This table contains daily time-series data and has 7,981,759 entries in total, covering the historical period from 1988-10-18 to 2016-10-21.

Table 2. Overview of database organization in terms of information groups, data tables, data fields, and total entries.

Information Group	Data Table	Data Fields	Category	Total Entries
Fund	Fund	Fund Identifier (FundID), Identifier in Morningstar’s Database (MorningstarID), Identifier in Fundsquare’s Database (FundsquareID), Identifier in the Commission de Surveillance du Secteur Financier (CSSF’s Database (CssfID), Legal Entity Identifier, Constitution Date, Status, and End Date	Life-Cycle	4591
	Fund.Profile	FundID, Name, Legal Form, and Fiscal Year End	Period-Based	4591
	Fund.Roles	FundID, Management Company, Is Self-Managed, Investment Manager, Administrator, Custodian, Auditor, and Transfer Agent	Period-Based	4591
	Fund.Contact	FundID, Street, City, and Country of Management Company	Period-Based	4591
Sub-fund	Subfund	Sub-fund Identifier (SubfundID), Identifier in Morningstar’s Database (MorningstarID), Identifier in Fundsquare’s Database (FundsquareID), Identifier in CSSF’s Database (CssfID), Legal Entity Identifier, Launch Date, Status, End Date, End Reason, Parent FundID, and PortfolioID	Life-Cycle	18,982
	Subfund.Profile	SubfundID, Sub-fund name, Currency, Pricing Frequency, Is Index Fund, Is Exchange Traded Fund (ETF), Is Fund of Funds, Is Master Fund, Is Feeder Fund, Master Fund ID, Is Leveraged, Is Socially Conscious, Is Sharia Compliant, Investment Area, and Prospectus Benchmark	Period-Based	18,982
	Subfund.Roles	SubfundID, Asset Manager, Investor Advisor, and Distributor	Period-Based	46,691
	Subfund.DailySeries	SubfundID, Date, Total Net Assets (TNA), and Currency	Time-Series (Daily)	7,981,759 (From 1988-10-18 to 2016-10-21)
	Subfund.ClassificationStatic	SubfundID, Date (Prospectus Date), and Morningstar Global Category	Period-Based	17,003
	Subfund.ClassificationDynamic	SubfundID, Morningstar Category, Morningstar Equity Style Box, and Morningstar Fixed Income Style Box	Time-Series (Monthly)	224,593 (From 1988-03-01 to 2016-10-31)
	Subfund.CashFlows	SubfundID, Date, Net Amount of Subscriptions, Net Amount of Redemptions, and Currency	Time-Series (Monthly)	984,673 (From 1988-03-01 to 2016-10-31)

Table 2. Cont.

Information Group	Data Table	Data Fields	Category	Total Entries
Portfolio	Portfolio.Summary	Portfolio Identifier (PortfolioID), Identifier in Morningstar's Database, Currency, and List of Report Dates	Life-Cycle	14,464
	Portfolio.Allocation	PortfolioID, Date, Sale Position, Number Holdings, Number Stock Holdings, Number Bond Holdings, Total Market Value, Asset Allocation Stock, Asset Allocation Bond, Stock Breakdown, and Bond Breakdown	Time-Series (Monthly)	250,445 (From 2002-10-31 to 2016-11-16)
	Portfolio.Holdings	PortfolioID, Date, Name, International Securities Identification Number (ISIN), Type, Sector, Country, Currency, Market Value, Weighting, Shares Held, Maturity Date, Coupon Rate, and First Bought Date	Time-Series (Monthly)	20,029,968 (From 2002-10-31 to 2016-11-16)
Share class	Shareclass	Share class Identifier (ShareClassID), Identifier in Morningstar's Database (MorningstarID), Identifier in Fundsquare's Database (FundsquareID), Identifier in CSSF's Database (CssfID), ISIN, Launch date, Parent SubfundID, Parent FundID, Status, End Date, End Reason, and Receiving Fund	Life-Cycle	84,556
	Shareclass.Profile	ShareClassID, Name, Currency, Distribution Flag, Is Currency Hedged, Hedged Currency, Is Institutional, Distribution Countries, and Minimum Investment	Period-Based	84,556
	Shareclass.DailySeries	ShareClassID, Date, TNA, Net Asset Value (NAV), Shares Outstanding, Bid Price, and Offer price Currency	Time-Series (Daily)	10,118,7743 (From 1988-03-01 to 2016-10-21)
	Shareclass.MonthlySeries	ShareClassID, Date, Return	Time-Series (Monthly)	4,657,460 (From 1988-03-01 to 2016-11-30)
	Shareclass.ClassificationStatic	ShareClassID, Morningstar Global Category, and Prospectus Date	Period-Based	14,337
	Shareclass.ClassificationDynamic	ShareClassID, Morningstar Category, Morningstar Equity Style Box, Morningstar Fixed Income Style Box, and Date	Time-Series (Monthly)	201,141 (From 1988-03-01 to 2016-10-31)
	Shareclass.FeesShareholders	ShareClassID, Subscription Fee Tier Structure, Redemption Fee Tier Structure, Switch Fee, and Prospectus Date	Period-Based	66,217
	Shareclass.FeesMaxima	ShareClassID, Max Management Fee, Max Ongoing Charge, Max Performance Fee, Total Expense Ratio, and Prospectus Date.	Period-Based	324,076
	Shareclass.FeesAnnualReport	ShareClassID, Management Fee, Ongoing Charge, Performance Fee, Total Expense Ratio, Turnover Ratio, and Annual Report Date	Time-Series (Annual)	128,821 (From 2000-09-25 to 2016-12-09)
	Shareclass.Dividends	ShareClassID, Ex-Dividend Date, Distribution Amount, Reinvestment Date, Payment Date, Declaration Date, Distribution Type, and Currency, Split Values	Time-Series	303,968 (From 1988-03-01 to 2016-12-20)

2.3. Historical Coverage

This section focuses on the historical coverage of the time-series fields. The distribution of the total number of entries (historical observations) across years is presented for the major time-series fields. More specifically, Table 3 presents the distribution of total entries from 1988 to 2016 for the following data tables: (i) Subfund.DailySeries, (ii) Shareclass.DailySeries, (iii) Shareclass.MonthlySeries, and (iv) Shareclass.Dividends. It is worth mentioning that the data table Shareclass.DailySeries is the largest table of the database considering the total number of historical observations.

Table 3. Distribution of the total number of historical observations per year for the major time-series.

Year	Total Number of Entries			
	Subfund. Daily Series	Shareclass. Daily Series	Shareclass. Monthly Series	Shareclass. Dividends
1988	17	35721	3013	71
1989	26	50,381	4211	121
1990	164	75,239	6237	143
1991	186	110,315	9024	219
1992	260	167,268	11,718	301
1993	373	234,591	15,049	373
1994	413	328,120	19,256	435
1995	498	423,602	23,153	556
1996	549	520,591	27,090	758
1997	700	659,915	32,737	899
1998	1286	860,097	40,846	1085
1999	4124	1,153,519	50,984	1414
2000	17,657	1,515,293	64,427	1635
2001	26,024	1,976,887	81,418	1967
2002	32,981	2,428,050	98,822	2651
2003	37,463	2,773,609	115,773	3662
2004	55,953	3,139,488	132,809	4593
2005	64,056	3,461,520	154,332	5203
2006	101,319	4,080,283	187,166	6444
2007	260,017	4,921,697	223,129	7780
2008	581,190	5,856,880	260,762	8549
2009	651,444	6,090,925	275,932	8418
2010	777,513	6,511,096	299,499	48,128
2011	790,429	7,137,134	331,239	42,191
2012	795,919	7,777,987	359,101	18,601
2013	877,118	8,689,339	396,759	24,974
2014	948,458	9,810,901	447,514	32,831
2015	1,044,755	10,955,075	497,283	39,273
2016	910,867	9,442,220	488,177	40,693
Total	7,981,759	101,187,743	4,657,460	303,968

2.4. Parent–Child Relationships

The structure of UCITS includes parent–child relationships among the information groups of fund, sub-fund, and share class because a fund may contain one or more sub-funds and a sub-fund may contain one or more share classes. The first type of parent–child relationship is represented in the database by linking the identifier of each parent fund with the identifiers of the underlying sub-funds. The second type of parent–child relationship is denoted by linking the identifier of each parent sub-fund with the identifiers of the underlying share classes. Figure 2 shows how the three core data tables (fund, sub-fund, and share class) are linked to each other, that is how the parent–child relationships of UCITS parts are represented in the database. In order to simplify the representation of the data tables, only the data fields related to identification information are presented in Figure 2.

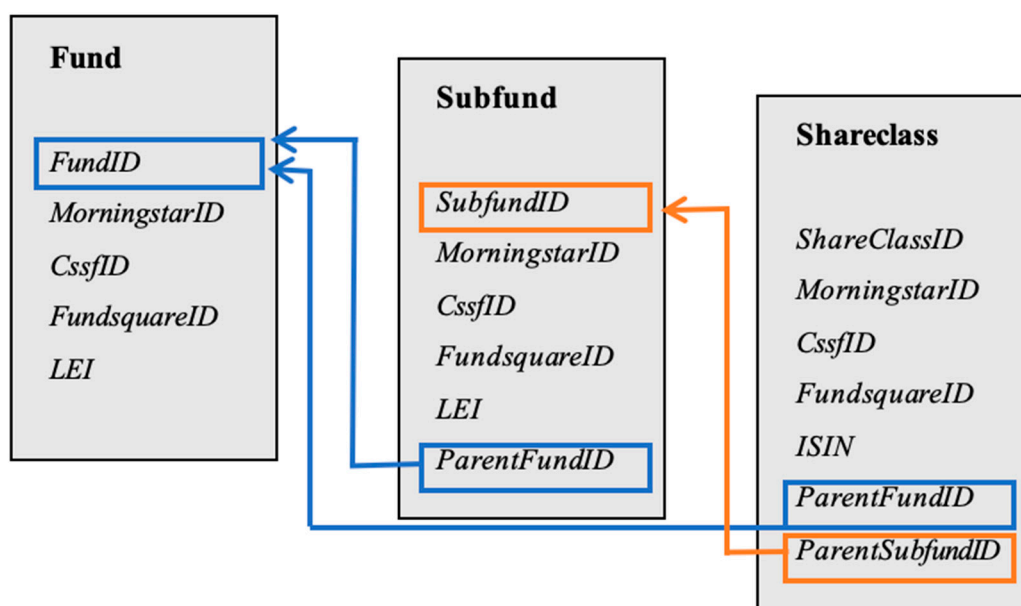


Figure 2. The three core data tables and the connections between them. The boxed data fields and their links point out the parent–child relationships.

3. Methods

This section describes the main steps of the database creation process with regards to the selection of data providers and the fusion of data collected from different sources. The contribution and the complementarity of the data providers is also presented. Finally, the process of entity matching and merging non-overlapping historical periods is described.

3.1. Selection of Data Providers

A set of data providers was specified in order to satisfy the primary objectives for the database content: the maximization of data field coverage, the maximization of fund population, and the maximization of historical coverage. Initially, a set of fields was specified after a thorough review of the existing investment fund literature and the current state-of-the-art research databases of investment funds. The European legislation that regulate this type of fund (UCITS European directives) and guidelines of the European Securities and Markets Authority (ESMA) were analyzed in order to determine UCITS-specific characteristics. Moreover, technical publications of the Association of the Luxembourg Fund Industry (ALFI) were studied as complementary information sources in order to clarify the structural particularities of UCITS that are domiciled in Luxembourg. The outcome of this step was a set of specific fields of UCITS on which the structure of the database was based, along with accurate definitions of the fields. Following that, an initial nine commercial providers of well-established financial datasets in the bibliography of investment fund research were investigated as potential data providers. The set of specified fields was the starting point for discussions with the potential data providers. Comparative statistics of the population of UCITS entities (namely the total number of funds, sub-funds, and share classes) were conducted among providers. Such comparative statistics are not disclosed in this paper due to restrictions of legal agreements. The criteria for the final decision of selection of data providers were in alignment with the three primary objectives of the database content: (i) the availability of the specified data fields, (ii) population availability among data providers, and (iii) the availability of historical observations for the specified data fields. Based on the maximization of these criteria, two commercial providers were selected, namely Morningstar and Fundsquare (the latter is a subsidiary of the Luxembourg Stock Exchange). In addition to these two commercial data providers, the CSSF kindly contributed to the construction of the new database by providing publicly available information from its historical archives. We used this contribution to

validate various aspects of the data against the data obtained from the data providers, in particular to validate the completeness of investment fund coverage in the datasets provided by the data providers. Finally, the available data of UCITS funds covering the period from 1988 to October 2016 were collected from Morningstar, Fundsquare, and the CSSF.

3.2. Complementarity of Provided Datasets

The provided datasets from Morningstar, Fundsquare, and the CSSF were analyzed in terms of population, data fields, and historical coverage. Table 4 presents an overview of the datasets provided and sheds light on the complementarity of the three data providers given the differences of population coverage and historical coverage between them. The first column of Table 4 (from right to left) organizes the data fields of the datasets provided in information groups (fund, sub-fund, portfolio, and share class), and the data fields are further classified into the categories of life-cycle fields and time-series data fields, as defined in the Section 3.2. The following three columns present the population (total number of entities) and the availability of historical data for each provider. For some fields, only the latest values of the fields were available in the providers' dataset as there exists an overwrite policy of the values. A characteristic example is the life-cycle fields of the fund level (such as the fund name or fund management company) in the datasets of Morningstar and Fundsquare, as both providers maintain only the latest value of these data fields. In contrast, the CSSF maintains and provided the complete history of such data fields from 1988 to 2016. As can be seen in Table 4, some of the fields were not available from providers. More specifically, Fundsquare could not provide any portfolio data because they do not maintain such data, and the CSSF could not deliver any portfolio data and time-series fields due to disclosure constraints. The final population of the LFDR after the process of data merging (described in the sub-section below) is presented in the last column of Table 4. Considering the information from the three data sources, the final database contains complementary data and achieves maximization with regards to population coverage and historical coverage.

Table 4. Population and history coverage of the provided datasets and of the Luxembourg Fund Data Repository (LFDR) content.

Information Group		Morningstar	Fundsquare	CSSF	LFDR
Fund	Life-cycle fields	Population: 1724 History: latest values only	Population: 2742 History: latest values only	Population: 4658 History: 1988–2016	Population: 4591 History: 1988–2016
	Life-cycle fields	Population: 17,728 History: latest values only	Population: 17,292 History: latest values only	Population: 19,289 History: 1988–2016	Population: 18,982 History: 1988–2016
Sub-fund	Time-series fields	Population: 9570 History: 1988–2016	Population: 14,965 History: 2000–2016	No data available	Population: 16,188 History: 1988–2016
	Time-series fields	Population: 14,464 History: 2002–2016	No data available	No data available	Population: 14,464 History: 2002–2016
Share Class	Life-cycle fields	Population: 75,045 History: latest values only	Population: 82,514 History: 1988–2016	Population: 88,462 History: 1988–2016	Population: 84,556 History: 1988–2016
	Time-series fields	Population: 69,549 History: 1988–2016	Population: 75,086 History: 2000–2016	No data available	Population: 81,056 History: 1988–2016

3.3. Matching Entities and Maximizing Historical Coverage

Given the three datasets of UCITS received from the different providers, it was important to correctly match the entities between the respective datasets in order to allow for the correct merging of datafields pertaining to identical entities. Ideally, entity matching would occur through a common

identifier for each fund, sub-fund, and share class. However, data providers typically use proprietary identifiers. Furthermore, the standard identifier, International Securities Identification Number (ISIN), is typically only to share classes—not to funds and sub-funds³.

A matching procedure had to be devised for each of the three levels (fund, sub-fund, and share class) for each UCITS provided by the data providers. It is worth mentioning that there was no need for the matching of portfolios because portfolio entities were only acquired by one provider (Morningstar). Thus, three types of entity matching were performed: fund, sub-fund, and share class matching. The decision criteria to accept or further investigate matches are described below. The matching of fund entities was based on string matching of the official fund names. Initially, an exact string matching algorithm was applied for the automatic matching of the fund names, and then a manual investigation of fund names took place for the automatically unmatched names. The matching of sub-fund entities was based on the two criteria: the matching of name of the parent fund and the matching of the ISIN identifiers of the sub-fund's underlying share classes. The matching of share class entities was based on the ISINs of share classes. Regarding the matching of entities (funds, sub-funds, and share classes), cases of unmatched entities existed. Unmatching was attributed to the misspelling of the names between data providers and/or the differentiation of ISINs. In such cases of ambiguity occurring in the matching procedure, the corresponding entities were not included in the final database. The latter decision was in accordance with the priority of ensuring the accuracy of the database.

In order to maximize the historical coverage for the time-series fields, the merging of historical observations was performed for the matched entities. As fund entities are not associated with time-series fields, the rest of this paragraph refers to time-series of sub-funds and time-series of share classes. For matched entities (either sub-funds or share classes), two cases were distinguished depending on the availability of time-series from one provider only or the availability of time-series from more than one provider. In the first case, the available time-series were introduced into the database for matched sub-funds and for matched share classes. In the second case, the merging of non-overlapping historical observations was considered given the two time-series in order to minimize the historical periods with missing data. A successful consistency check was required before merging non-overlapping historical observations for matched entities. The consistency check focused on a set of five common dates between the two time-series to be merged, and the check was successful when the values of the time-series coincided. Such consistency checks can be considered as an additional validation check for matched entities (matched sub-funds and matched share classes). After the merger of historical time-series, the resulting database contained time-series of higher historical coverage for sub-funds and share classes compared to the corresponding datasets received from the original providers.

4. User Notes

This section discusses potential academic applications of the database, the expected impact of the database, and its future development plans.

The database brings innovation to academic research in the investment funds by establishing a dataset that is unique in terms of both its structure and content. Historical data available in the Luxembourg fund industry heretofore were fragmented, incomplete, and unsuitable for academic research. The dataset provides a functioning and qualified platform for the expansion of academic research beyond the population of US mutual funds into an alternative population of investment funds with a much greater degree of global reach in terms of investors and investment portfolios. Such a tool was not available to researchers before the construction of the LFDR. The database is a unique and complete fund dataset from the largest UCITS fund domicile, Luxembourg. This population

³ In the absence of an assigned sub-fund ISIN, market practice typically requires the ISIN of the main share class to identify a sub-fund portfolio for the purposes of processing its trade settlement instructions. For the purposes of the settlement of subscription and redemption payments from and to the fund shareholder, the relevant share class ISIN is used. Thus, only share classes have ISINs in use in market operations.

of funds can be said to be much more representative of global investment activity than US mutual funds. The unique fund domicile neutralizes the factor of potential differences in regulation that would stem from a population comprising funds from different domiciles. Moreover, it allowed for the construction design of the database that was tailored to the operational and data specificities of funds in this domicile, and it allows for a mechanism to ensure that the dataset can be verified for completeness. Through the inclusion of funds that have been closed or merged into other funds, it ensures the absence of survivorship bias, which is an importance feature for datasets of academic research [11].

In all the above respects, the innovating characteristics of the database make it a tool for a population of globally representative investment funds that is equivalent and comparable to the CRSP Mutual Fund database for its population of US mutual funds [12]. However, the LFDR provides an important additional innovation: the inclusion in the dataset of investment flows into and out of each fund portfolio and share class. Monthly data on investment flows at the sub-fund level will provide substantial scope for research into the subject population that the CRSP does not currently provide for its population: investment flows in their own right, as well as their (relative) correlations with specific fund product characteristics, environmental variables (such as economic conditions, market news, and other one-time events), and other external phenomena.

Publicly sold investment funds such as UCITS and US mutual funds are highly regulated, providing a compact and readily observable set of data on many dimensions. Thus, the immediate impact of the database will be to allow for the pursuit of lines of research for the Luxembourg UCITS fund set that have been established for many decades for US mutual funds and that have developed into a rich body of literature. The extant literature on US mutual funds can be said to fall into three broad categories: underlying portfolio preferences of investors or investor segments in terms of asset allocation, portfolio strategies, and investment trends; the science of portfolio management in terms of explanatory and predictive models with respect to portfolio risk and return optimization; and measures of funds as investment products in their own right, such as comparative performance parameters, across funds and individual funds' performance persistence over time, and the effects their taxation and regulatory regimes. In addition, the investment funds flow data of the LFDR will allow for research in investment funds along entirely new lines that have not been possible for US mutual funds in the absence of flow data in the CRSP Mutual Fund database.

The future development of the database is currently being discussed by the project team at the University of Luxembourg. As the current dataset contains population of UCITS until 2016, the future plans focus primarily on the integration of the data from 2016 to now, as well as the provision of regular updates in the future. In addition, making the database available to the global academic community under license restrictions is being investigated and would require appropriate agreements with the data providers. Other future work includes a search for still missing historical data and the resolution of the remaining identification issues resulting from unmatched fund and sub-fund entities. Regarding the still missing historical portfolio data for some funds, the option of soliciting additional data providers is being considered. Finally, error reporting and correction mechanisms are being developed to identify, validate, and process any errors reported by the users of the database.

Author Contributions: J.P. and J.S. conceived the study. A.S. and J.P. designed the study, and they collected and analyzed the data. A.S. designed and implemented all technical parts of the study and of the database creation. A.S. wrote an initial draft on the basis of the database. J.P. and J.S. critically revised the draft manuscript and made important changes to the content. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ALFI and the LSF Foundation.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

This section clarifies the main characteristics of the investment fund type of UCITS. The following paragraphs correspondingly present the three structural levels of UCITS, namely the fund level, the sub-fund level (including portfolio information), and the share class level. The hierarchy among

these levels and the characteristic information per level are illustrated in Figure A1. A simplified example of UCTIS is also presented in Figure A1 to illustrate the structure with respect to funds, sub-funds, and share classes. This UCITS example is a simplified part of an actual fund and is used only here for illustration purposes. The information of this example is publicly available [13].

UCITS constitute a legal entity formed either as a single fund or as an umbrella fund consisting of multiple sub-funds (also known as compartments) [14]. Though an umbrella fund consists of a number of sub-funds, it forms a single legal entity. Considering that the vast majority of UCITS are umbrella funds and that a single-compartment fund can be seen as a simplified version of an umbrella fund, the rest of this section assumes an umbrella fund model. UCITS may have appointed a management company or may be part of a self-managed UCITS investment company. In Luxembourg, the UCITS management companies are authorized according to Article 101 of Chapter 15 of the 2010 Luxembourg Law related to undertakings for collective investment [15]. Apart from the management company, other service providers related to the fund level of UCITS include services of transfer agency, fund accounting, and custodian/depositary services. UCITS that are domiciled in Luxembourg must be authorized by the CSSF before beginning their activity. Afterwards, they are supervised by the CSSF on an ongoing basis by means of regular reporting. The disclosure requirements of UCITS include fund prospectuses, extensive annual reports, and the publication of diverse monthly and semi-annual time-series.

According to the work of [14] and in accordance with Article 29(1) of the EU Regulation No 1095/2010 of the European Parliament, sub-funds are separate parts of a common fund vehicle and have their own investment objectives. Assets of one compartment, also known as asset portfolio or pool of assets, are distinct from assets of other compartments. Compartments are usually legally segregated from other compartments, meaning that a liability arising in one compartment cannot be offset by the assets in other compartments of the fund. In other words, each sub-fund corresponds to a distinct asset portfolio and distinct liabilities. In addition, each sub-fund differs in its investment strategy from the other compartments and may appoint its own investment manager and/or investment advisor apart from the investment manager of the entire fund [9].

While the European directives on UCITS cover funds and compartments, they do not clearly present the definition and scope of share classes, although they recognize their existence [15]. In fact, UCITS or one of their compartments can be sub-divided by share classes. Share classes are categories of share that belong to the same UCITS and allow for subsets of investors in UCITS to achieve some level of customization. Such customization accommodates the specific needs of the investors or the terms and conditions for the subscription into the fund (e.g., a distinct fee structure, the distribution or capitalization of revenues, a particular tax treatment under national law, or a distinct minimum investment amount) [14]. The share classes in a sub-fund together have the same claim to a single pool of assets (namely the asset portfolio of the sub-fund) and there is no segregation of these assets between share classes⁴. The asset value of each share class is determined by an apportionment of the change in value of the pool of assets on the basis of a distribution coefficient. Though there is no legal segregation of assets between share classes, expenses defined as pertaining to specific share class are attributed to that share class only and are integrated into the calculation of its share value. Any investment outcome relating to specific arrangements for a given share class, such as a foreign exchange hedge transaction, is attributed to that share class only. In terms of information disclosure, the CSSF requirements include the preparation of a key investor information document for each share class.

⁴ A possible exception is an asset, such as a currency receivable, representing claims related to foreign exchange hedging transactions undertaken exclusively on behalf of a specific share class. In such a case, the asset (and its fluctuating value) insures solely to the share class and is incorporated only into that share class's share value.

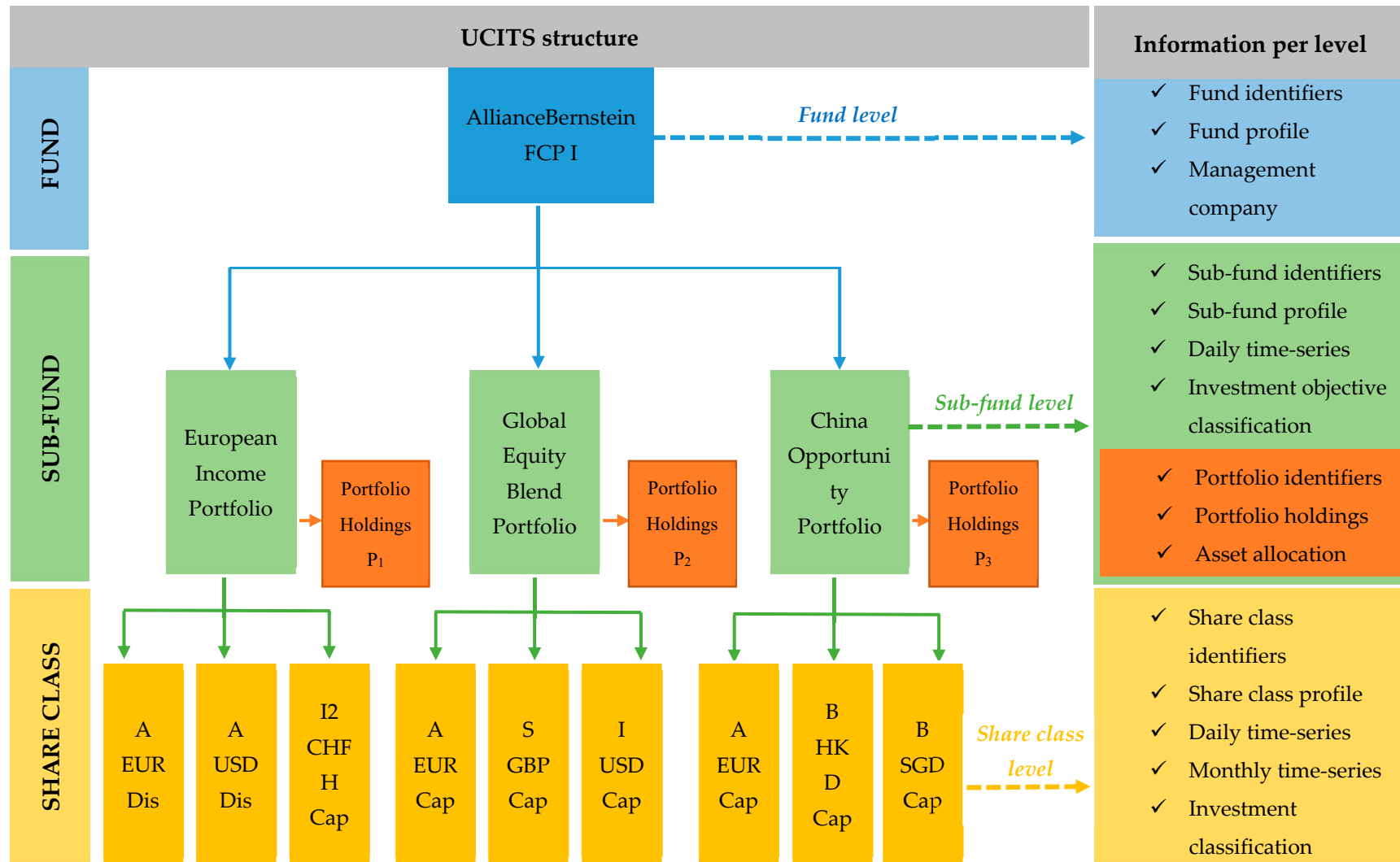


Figure A1. Structure of an umbrella Undertakings for Collective Investment in Transferable Securities (UCITS) fund with characteristic per level information.

Appendix B

This section presents the full text of the license agreement that the researchers of the Department of Finance at the University of Luxembourg need to sign before accessing and using the database content. (Figure A2). The license agreement presents the legal constraints resulted from the contractual agreements signed with the data providers (namely Morningstar, Fundsquare, and the CSSF). The main restrictions imposed by the license agreement are the use of the data for academic research purposes only (i.e., not commercial purposes) and the obligation of not transferring the database content outside the University of Luxembourg.

<u>License agreement of the Luxembourg Fund Data Repository</u>
<p>I, the undersigned, _____, acknowledge the following terms of use for the above database. The term “data” used herein refers to the data in the above database and to any part of it.</p> <ul style="list-style-type: none"> • The undersigned acknowledges that he/she has requested permission to use the above data for his/her professional research purposes and that the University has consented to this request on this basis under the restrictions enumerated below. • The undersigned acknowledges that the data is intended for professional research purposes only for researchers employed by the University of Luxembourg. • The undersigned acknowledges that the data has been provided to the University of Luxembourg by data providers for use by University of Luxembourg researchers under their employment with the University, and that these have imposed strict contractual conditions on its use and its disposition including the transfer of the data to other parties. • The data may not be shared with, transferred to, made available or rendered accessible in any way to other parties inside or outside the University of Luxembourg. • The data may not be copied onto a computer that is not University of Luxembourg equipment assigned to the undersigned. • If the undersigned has finished the research project, is obliged to surrender the computer on which the data is stored to the University, is no longer employed by the University, or if the computer on which the data is stored becomes the property of a party other than the University, the Undersigned undertakes to permanently delete the data remaining on the computer. • The data remains the property of the University of Luxembourg at all times. <p>Acknowledged by: _____ Signature _____ Date _____</p>

Figure A2. The full text of the license agreement.

References

1. *Investment Company Institute Fact Book 2019: A Review of Trends and Activities in the Investment Company Industry*, 59th ed.; Investment Company Institute: Washington, DC, USA, 2019; pp. 26–30.
2. Delbecque, B.; Tilley, T.; Yang, H. *Trends in the European Investment Fund Industry in the Fourth Quarter of 2019 & Results for the Full Year of 2019*; European Fund and Asset Management Association: Brussels, Belgium, 2019; pp. 1–6.

3. *Annual Report 2018–2019*; Association of the Luxembourg Fund Industry, ALFI publications: Luxembourg, 2019; pp. 3–12.
4. Wermers, R.; Smith, R. Are Mutual Fund Shareholders Compensated for Active Management «bets»? Available online: <http://terpconnect.umd.edu/~wermers/FAJ%20Paper%20Post-Submitted%20Version.pdf> (accessed on 18 May 2020).
5. Carhart, M. On Persistence in Mutual Fund Performance. *J. Financ.* **1997**, *52*, 57–82. [CrossRef]
6. *Setting up in Luxembourg*; Association of the Luxembourg Fund Industry, ALFI publications: Luxembourg, 2019; pp. 1–3.
7. Chambost, I.; Lenglet, M.; Tadjeddine, Y. *The Making of Finance: Perspectives from the Social Sciences*, 1st ed.; Routledge: London, UK, 2018.
8. European Commission: EU Laws and Initiatives Relating to Collective Investment Funds. Available online: https://ec.europa.eu/info/business-economy-euro/growth-and-investment/investment-funds_en (accessed on 18 May 2020).
9. Luxembourg Stock Exchange: UCITS. Available online: <https://www.bourse.lu/ucits> (accessed on 18 May 2020).
10. *The UCITS Market Geographical Breakdown of Nationally Domiciled Funds*; European Fund and Asset Management Association: Brussels, Belgium, 2019; pp. 1–2.
11. Elton, J.E.; Gruber, M.J.; Blake, C.R. Survivor Bias and Mutual Fund Performance. *Rev. Financ. Stud.* **1996**, *9*, 1097–1120. [CrossRef]
12. Documentation of CRSP Mutual Fund Database. Available online: <http://www.crsp.org/products/documentation/crsp-survivor-bias-free-us-mutual-fund-guide-crpsift> (accessed on 18 May 2020).
13. Fund information. Available online: <https://www.fundsquare.net/fund-tree?idInstr=154601> (accessed on 18 May 2020).
14. *Discussion Paper on UCITS Share Classes*; European Securities and Markets Authority: Paris, France, 2016; pp. 3–6.
15. Law of 17 December 2010 Relating to Undertakings for Collective Investment. Available online: https://www.cssf.lu/wp-content/uploads/FAQ_Law_17_December_2010_100320.pdf (accessed on 18 May 2020).



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).