Ref. Ares(2025)5353012 - 03/07/2025

TWINRD
ECODIGITAL FUTURES

MACROECONOMIC MODELLING OF R&D FOR THE TWIN TRANSITION

D2.1 | Technology flows matrices for green and digital technologies

Author(s)	**Önder Nomaler (UNU-MERIT), Bart Verspagen (Maastricht University)**
Version	**2.0**
Quality review	
Date	
Dissemination level	**SEN**
Grant Agreement	**N° 101132170**
Starting Date	**01-01-2024**
Duration	**36 months**
Coordinator	**Leonidas Paroussos (E3-Modelling)**
E-mail	**paroussos@e3modelling.com**

Funded by the European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or CINEA. Neither the European Union nor the granting authority can be held responsible for them.

Document History

Date	Contributors	Action	Status
2 November 2024	Önder Nomaler (UNU-MERIT), Bart Verspagen (Maastricht University	Dataset and documentation for green and digital patent counts without further breakdown shared on project's sharepoint	Draft (V1.0)
11 December 2024	Önder Nomaler (UNU-MERIT), Bart Verspagen (Maastricht University	Dataset and documentation for green and digital patent counts with further breakdown shared on project's sharepoint	Draft (V1.1)
3 July 2025	Önder Nomaler (UNU-MERIT), Bart Verspagen (Maastricht University	Dataset and documentation for technology flow matrices added and overall document edited and finalized	Final (V2.0)

Abstract / Executive Summary:

This document explains the construction and usage of the TWINRD datasets on patent counts and knowledge flows. It is an accompanying document to an online dataset (links will be given below), which is based on the PatStat database by the European Patent Office. We processed the PatStat data into patent counts and patent citation-based (knowledge) flows indicators that will be used in other parts of the project. Because of the size of the three separate datasets that we supply, data are provided in a tabular format that needs further data extraction, such as the construction of a pivot table, or data queries for an SQL-based language, or software like Python or R. We provide a set of example SQL queries that illustrate the working of these data extraction principles from the data that we supply. Users of the data (either counts or knowledge flows) are encouraged to read the entire document, as it describes the principles on which we built the data, which are crucial to understand for proper use of the data.

1. Introduction

This is a technical document describing the construction of a dataset of patent counts and technology flows that are to be used in the TWINRD project. The document also describes how the dataset is to be used, i.e., how meaningful data on technology flows can be extracted from the raw data as provided. Although, in a strict sense, the deliverable itself consists of the technology flows data only, this document also describes the development and use of the patent counts that underlie the technology flow matrices. These patents counts were also used in other parts of the project, for example, in Deliverable 2.2 ("R&D data for energy, green and digital technologies"). This document provides the basic background information for all work on patent data in the project.

The technology flow matrices are based on patent citation data. The basic idea is that a cited patent provides some flow of knowledge that is used in the citing patent. The basic principle on which the technology flow matrices are built is therefore to count the number of times that patents that are invented in one sector in one country are cited in a(nother) sector in some (other) country. We provide data for the entire set of countries and (NACE) sectors found in the PatStat database as supplied by the European Patent Office (EPO), which is our source data for all data that we produce. The breakdown that we provide for countries and sectors is the most detailed one possible, and users of the data will likely want to aggregate this into groups of countries and / or more aggregated sectors. The document provides guidelines on how to perform this aggregation.

As the TWINRD project is concerned with both "green" and "digital" technologies, the identification of patents in each of these groups is a major part of the efforts undertaken. For patent counts, we provide two datasets: one where green and digital patents are aggregated into all green and all digital (although we use two different definitions of digital), and one in which green and digital are further broken down into sub-categories.

For the technology flow matrices, we only provide data for all green and all digital. Although we are, in principle, able to apply technology flow matrices containing a more detailed breakdown of green and digital, in the project as a whole there is currently no decision on what the best breakdown will be. To provide technology flow matrices using the fully detailed breakdown of green and digital that was used in the counts would be unpractical, because this will lead to a very large and fragmented dataset (many combinations of cited and citing, countries, sectors, and green and digital sub-categories, many of which will contain very small numbers of patents). Once the project takes a decision about the desired level of green and digital breakdown in the technology flow matrices (if such a breakdown is desired at all), we will update the database.

This document assumes that the reader has some degree of familiarity with patent data and/or the PatStat database. We will use information on country of origin of a patent (by inventor), on the patent office where a patent was applied for (including international offices), application date of the patent, and technological classification systems of patents (such as IPC and CPC). We will also rely heavily on the concept of a patent family (in particular the DocDB families in PatStat). A DocDB patent family is a set of individual patent applications that are supposed to describe a single invention, for example different patents of the same invention that seek protection in different jurisdictions. DocDB families are our basic unit of counting, both in the datasets on patent counts and on technology flows.

The rest of this document is organized as follows. In the next section, we will describe how patent counts were performed, how this procedure is reflected in the basic dataset on counts, and how meaningful aggregated data can be extracted from the basic dataset. This will include an

explanation about the definition of green and digital technologies. In Section 3, this is extended to a breakdown of digital and green technologies into sub-categories. Finally, Section 4 explains the construction of the technology flow matrices.

2. Fractional patent family counts in green, digital and other technologies

In the 2023 autumn edition of the EPO PatStat database, we consider all DocDB patent families where the earliest year of the applications in the family is 2000 or later. We identify about 50 million families in this way. On the basis of the underlying applications ("family members"), we break these families down to combinations of countries and industries by fractional counting.

For countries we consider the inventors' (as opposed to the applicants') country of origin as indicated in the address in the patent application. In terms of industries, we use the fractional PatStat mapping (i.e., table tls229_appln_nace2) which assigns one or several NACE2 codes to each patent application. The assignment is based on a 'one-to-one' concordance between (a combination of 4 and 8 digit) IPC codes and 84 NACE codes at 2-digit level. This concordance is provided within PatStat, and has been constructed on the basis of the most likely industry of origin of the patent (e.g., a plough would be assigned to the agricultural machinery sector rather than to agriculture, where it would most likely be used). Since patents are typically assigned more than one IPC code, PatStat can assign more than one NACE2 code to a patent application, in case of which the presence of NACE2 will be indicated by a weight.

An example will illustrate the principle of fractional counting of patents over countries and sectors best. We start at the level of individual patent applications, i.e., patent family members. Say, a patent application has 5 inventors, 4 from France and 1 from the Netherlands. In our fractional counting, that would imply a 0.8 contribution to France and 0.2 to the Netherlands. Furthermore, say, this patent application is assigned to the NACE2 codes 13 (manufacturing of furniture) and 15 (Manufacture of Leather and Related Products) with the respective weights of 0.6 and 0.4. In our fractional counting method, this patent application will be considered as a contribution to the French furniture industry for $0.8 \times 0.6 = 0.48$, to the Dutch furniture industry for $0.2 \times 0.6 = 0.12$, to French leather manufacturing for $0.8 \times 0.4 = 0.32$ and to Dutch leather manufacturing for $0.2 \times 0.4 = 0.08$. Obviously, the assignments add up to 1 patent application.

In a next step, we add these contributions up over family members for each patent family, and divide each contribution by the number of applications in the family. In this way, we make sure that all the country-industry combinations that are present under the family add up to 1 (family).

Next, we identify whether a patent family can be associated to green-tech and/or digital-tech. For green-tech, we adopt the OECD EnvTech definition which provides a list of CPC codes under various sub-categories. These CPC codes largely draw on EPO's Y02A-Y04S tagging scheme that identifies the technologies for the mitigation of and adaptation to climate change, but also broadens the definition with additional categories such as air/water pollution abatement, waste management, soil remediation etc. We consider a given patent family as 'green' if at least one of its applications has at least one CPC code that can be found in the OECD EnvTech list.

For digital technologies we use two different definitions in order to differentiate between the recently emerging ('new') digital technologies (e.g., 4IR) and the established ('earlier') ones. For the latter ('earlier'), we draw on OECD's J-Tag scheme, which is essentially a list of IPC codes classified under various sub-categories of digital technologies (such as mobile communications, electronic measurement etc.). Similar to the case of green-tech, we label each patent family in a

binary fashion in view of its 'early-digital' (= J-tag) content according to whether at least one of its applications has at least one IPC code that can be found in the OECD J-tag list.

For the 'new' (emerging) digital technologies we draw on the 2022 WIPO World Intellectual Property Report 2022: The Direction of Innovation, which provides a search strategy to identify emerging digital technologies (which are referred to as digital 'General Purpose Technologies') under several categories such as artificial intelligence/machine learning (AI/ML), Internet of things (IoT), autonomous systems, big data, or robots. Unlike EnvTech and J-tagging, the WIPO patent search strategy contains, in addition to a list of CPC codes, also keywords and keyword-CPC combinations. We use an updated version of this search strategy that was kindly made available by the WIPO office in the form of a database query.

We implemented the WIPO query in PatStat, which required some adaptation to our MS SQL dialect. Our query identified about 2.5 million patent applications as related to a digital GPTs. We label each post-2000 patent family in PatStat in a binary fashion as WIPO-digital if it has at least one member that can be found in this list of 2.5 million applications.

All these pieces of information were aggregated into a dataset of fractional patent family counts. This dataset takes the form of a database table, where each row describes one subset of patents as defined by the characteristics that we discussed so far, and gives the patent family count for these characteristics. The following example shows three rows from this table:

earliest_filing_year	Inv_Country	nace2_code	has_Green	has_JTag	has_WIPO_Digtl	fr_nr_families
2020	CN	26.3	0	1	0	16205.07561
2006	JP	26.1	1	1	0	356.6453582
2010	US	26.2	1	1	1	44.4451684

The first row is interpretable as follows: in 2020, 16,205.076 patent families that were not green, not digital as defined by our WIPO-digital classification, but were green as defined in the J-tag classification scheme, were invented in China (CN) in the Communication Equipment Manufacturing sector (NACE 26.3). For the second row, we see that in 2006, 356.645 patent families that are green and digital in the J-tag scheme, but not digital in the WIPI scheme, were invented in Japan (JP) in the Electronic Components and Boards Manufacturing sector (NACE 26.1). Finally, the third row shows that in year 2010, 44.445 patent families that were green, digital in the J-tag scheme and digital in the WIPO scheme were invented in the US in the Computers and peripheral equipment manufacturing sector (NACE 26.2).

These data are available on the Zenodo platform as described in the Box. This takes the form of an MS EXCEL file that contains the full version of the above example table (with all existing combinations given in 253,362 separate rows), as well as information on classifications and a pivot table example for extracting data.

This concludes the description of the first of the two patent count datasets. There are, however, a few caveats to the procedure that we must report. First, due to the imperfection of administrative practices at national patent offices (especially China and South Korea), for a significant number of patent applications, the inventor address including national origin is not reported in PatStat. For applications made at international patent offices, there is nothing one can do to remedy this (in our fractional accounting we assign "NF", as in 'not found' to such cases). However, for applications made at national offices, we assume that the inventors without address (country) originate at the country where the application was made.

Second, the industrial (NACE2) codes are not available for all patent applications in PatStat. The main reason is that (as indicated above) the assignment of NACE2 codes to patent applications are based on a correspondence between IPC and NACE2 codes and quite a few patent offices use the CPC scheme instead of IPC. The CPC taxonomy is significantly finer-grained than IPC and it is not straightforward (for EPO, as the producer of the PatStat database) to map all CPC codes to IPC. Although we account for the missing NACE2 codes in our fractional counting practices (i.e., missing NACE2 codes lead to summed NACE2 weights that are less than 1, and hence do not get included in the counts), the dataset we provide only indicates the cases where the NACE2 codes are available in PatStat.

Third and finally, our patent family counts implicitly assume that the family is a proper 'unit' of knowledge produced by an invention. However, not all inventions are equally valuable (technologically and/or economically). The literature on the economics of innovation suggests that patent families with only one application made at a national office are less valuable than those with applications made at multiple national offices and/or at international offices such as the EPO.

This gives us a wide range of alternatives for quality/value assessment. One well-known example is the notion of a 'triadic-family' (i.e., a family which has at least applications made at EPO, USPTO and the JPO). Given the fact that the focus of our project is on the future potential of new technologies (be it green or digital), and the emerging role of China and South Korea in these technologies is large, we believe that the triadic-family definition is too restrictive. Thus, we adopt a different, less stringent filtering strategy. We define our 'quality filter' as having at least one application made at an international office (such as EPO or WIPO), or having applications made at least at two different national offices.

The inevitable, but necessary implication are drastic: out of the 50+ million post-1999 patent families one can find in PatStat, our quality filter leaves us with about 6.3 million patent families. An interesting consequence is that the quality filtering increases the presence of digital by about 15 percentage points at the expense of technologies that are neither digital nor green. A comparative breakdown of the filtered and non-filtered family counts regarding all 8 possible combinations of out green, old-digital, new-digital categories are provided in the last sheet of the main MS EXCEL data file in the Zenodo download (see Box). Unfiltered counts can be obtained in the second dataset that will be explained in the next section.

Box – Access to the data on Zenodo

The datasets as explained in this document are available for download from the Zenodo platform. The link for the first dataset is https://zenodo.org/records/15776235

This contains two Excel files. The file named

AllTech_Fractional_FamilyCounts_From2000_Onwards_Selected_Filter.xlsx

Has the main data in the sheet called *Frctnl_Family_Counts_Filtered*. There are additional sheets with a pre-formatted Pivot table, the NACE2- and country-codes, and a sheet called *Summary_All_Years_TechComb* that provides a comparison of aggregates with and without quality filtering. Data in the sheet *Frctnl_Family_Counts_Filtered* are quality-filtered as described in the text (Section 2).

The file named *Codes_NACE_Country_TechBreakdown.xlsx* contains the country codes, NACE2 sector codes and a breakdown of the J-tag, WIPO-digital and EnvTech classification systems. In terms of the first dataset, these last breakdowns are mostly relevant as a reference to the content of each of these classification systems, because this dataset only contains for aggregated versions of these classification systems.

The link for the second dataset is https://zenodo.org/records/15776253

In this dataset, there are two zipped csv files. The first of these is named

DetailedTech_Fractional_FamilyCounts_From2000_Onwards_FilterIs_NoFilter.zip or *.csv*

and contains the unfiltered patent counts broken down by detailed green (EnvTech) and digital (WIPO and J-tag) classes. The file named

DetailedTech_Fractional_FamilyCounts_From2000_Onwards_FilterIs_GT0_Internatnl_Or_GT1_Natnl.zip or *.csv*

contains the filtered version of the second dataset (as explained in the text, Section 3).

Finally, the link for the third dataset (knowledge flows) is https://zenodo.org/records/15796025

In this dataset, there are two zipped csv files named

Fractional_Citations_6_yearlyCohorts_Of_CitedFam.zip or *.csv* and
Fractional_Citations_6_yearlyCohorts_Of_CitingFam.zip or *.csv*

of which the first contains the 'forward-looking' knowledge flows dataset, and the second contains the 'backward-looking' dataset, both as explained in the text (Section 4).

3. Fractional Patent family counts in detailed technological subcategories of Green and Digital Technologies

As explained previously, we identify and tag patents (binarily) on whether they are EnvTech (i.e., green), and/or J-tag-digital (i.e., the 'early' digital) and/or WIPO-digital-GPT (i.e., as the 'new' digital). Even with this 3-way binary tagging approach, we already had $2^3 = 8$ possible technology profiles for each family: In terms of the variable names used in the dataset, the binary vector [has_Green, has_JTag, has_WIPO_Digtl] can show 8 different configurations such as [0,0,0], [0,0,1], [0,1,1], … , [1,1,1] for any given patent family. This adds to the multiplicity of inventor countries and NACE2 industries at the patent level.

Each of the EnvTech, J-tag and WIPO-digital tags can be broken down into sub-categories. Without going into details about how the sub-categories are constructed, we note that WIPO-digital features 10 subcategories (such as AI/ML, IoT, robots, blockchain, big data, cloud computing…etc.), while that of the OECD J-tag scheme has 13 (e.g., high-speed network, mobile communication, security, human-interface…, etc.), and OECD EnvTech 14 subcategories (Y02A, Y02B…, etc.). That makes, in total, 37 subcategories to profile each patent family separately in a binary fashion.

A major issue of concern here is the fact that patents typically have more than one IPC/CPC code, and also their titles/abstracts are a collection of many words. Thus, it is often the case that a given patent family qualifies in more than one technological subcategory. For instance, a given patent family that is broadly tagged as has_WIPO_Digtl=1 can be tagged simultaneously in IoT, Robots and Cloud Computing among the 10 WIPO-digital categories, and furthermore both in Y02A and Y02P among the 14 green subcategories of EnvTech. Considering the 10+13+14 = 37 categories we have in total, there are 2^{37} possible ways in which a given patent family could be profiled in terms of binary strings of the 37 sub-categories of WIPO-digital, J-tag and EnvTech.

Considering this potentially-overwhelming number of patent profiles, we chose not to account for the sub-technology detail in the first version of the dataset, i.e., in the dataset as described in the previous section. However, we provide also a second dataset in which we do provide full detail on the sub-categories. Here we observe that, out of the 2^{37} possible WIPO-digital, J-tag and EnvTech combinations, only around 21,000 combinations have been realized in reality. This results in a dataset with about 1.9 million rows (where the quality/value controls explained earlier have been imposed).

The new (second) dataset is also available on Zenodo (see Box for details). This also contains a version of the second dataset without any quality/value restrictions (which has now about 2.5 million rows, with about 34,000 sub-technology profiles). As we will explain below, the unfiltered second dataset can also be used to construct the unfiltered first dataset.

The new dataset is structured in the same way as the first dataset (previous section). Before, each row had a combination of country, NACE2, green (yes or no), J-tag (yes or no) and WIPO-digital (yes or no). Because J-tag-, EnvTech and WIPO-digital are now split up into a total of 37 sub-categories, we have both more rows and columns in this second dataset. We also introduce an additional dimension to our fractional counting method, which will be discussed below. When aggregated to just yes/no for J-tag, WIPO digital and EnvTech, the new dataset will produce identical results to the previous one.

In the first dataset, it was relatively easy to calculate aggregates, e.g., all patents for a country and NACE2 sector, irrespective of whether they are green or digital. Such an aggregation could, for example, be done by MS EXCEL's pivot table tool. Due to the sheer size of the new dataset (e.g., it does not fit into Excel), any desired aggregations will need to be performed on a different software platform, such as a database engine (with SQL scripting) or Phyton or R. In the rest of this document, we will provide examples of various aggregations in standard SQL scripts. It is straightforward to translate these into Phyton or R (whether or not with the help of AI!).

In the remainder of this section, we will illustrate aggregation in our second dataset with the help of SQL scripts, where we assume that the dataset is loaded into a database table named *Fractional_Counts_Detailed*. In our first example, we simply aggregate the data over all years, to find out how many (fractional) patent families are held by each NACE2 industry of each country, regardless of being digital or green, and regardless of timing.

```
SELECT Inv_Country, nace2_code, SUM(fr_nr_families) AS total_fr_nr_families
FROM Fractional_Counts_Detailed
GROUP BY Inv_Country, nace2_code
```

This will lead to an aggregated dataset in long format (with about 11.5k rows) which can as well be transformed to a wide-format (i.e., a matrix with countries on the rows and NACE2 categories on the columns).

If we are only interested in patenting in GreenTech (unconditionally on being WIPO-digital and/or J-Tag-digital). The following SQL script will do the job.

```
SELECT Inv_Country, nace2_code, SUM(fr_nr_families) AS total_fr_nr_families
FROM Fractional_Counts_Detailed
WHERE has_Green = 1
GROUP BY Inv_Country, nace2_code
```

In order to illustrate the convenience of fractional counting with the 2^3=8-way technological binary split of J-tag, WIPO-digital and EnvTech, we consider an even simpler example where we consider only French patents, overlooking momentarily the NACE2 detail:

```
SELECT SUM(fr_nr_families) AS total_fr_nr_families
FROM Fractional_Counts_Detailed
WHERE Inv_Country = 'FR'
```

This query will provide us with the scalar value: 203935.648935.

Let us further break this number down to two groups: green and non-green patents.

```
SELECT has_Green , SUM(fr_nr_families) AS total_fr_nr_families
FROM Fractional_Counts_Detailed
WHERE Inv_Country = 'FR'
GROUP BY has_Green
```

which will provide us with the following table

has_Green	total_fr_nr_families
0	178465.5547
1	25470.09425

Observe that the numbers in the second column add up perfectly to the total number of French patent families that was computed by the previous SQL query.

We can break these figures further down, also considering the digital content (say, in the sense of the WIPO-digital), which will yield the 2^2=4 possible combinations of green and WIPO-digital.

```
SELECT has_Green , has_WIPO_Digtl, SUM(fr_nr_families) AS total_fr_nr_families
FROM Fractional_Counts_Detailed
WHERE Inv_Country = 'FR'
GROUP BY has_Green, has_WIPO_Digtl
```

This will yield the following table:

has_Green	has_WIPO_Digtl	total_fr_nr_families
0	0	172854.0106
0	1	5611.544098
1	0	25120.07504
1	1	350.0192046

Once again, observe that not only the (fractional) figures on the third column perfectly add up to the total number of the French patent families, but those on the top (bottom) two rows perfectly add up the first (second) row of the second column of the table produced by the previous query.

What we have demonstrated with the SQL queries so far also reflects how the first version of our dataset worked. Now let us explore the possibilities that come with the new dataset which breaks each of our earlier 3 technological categories (i.e., captured by the earlier binary variables has_Green, has_JTag, has_WIPO_Digtl) further down to their respective subcategories.

As mentioned, under the main GreenTech (i.e., OECD EnvTech) category, we have 14 subcategories. Each patent family is accordingly tagged as a combination of the following set of 14 binary variables:

VarSet_EnvTech = [*has_Y02A, has_Y02B, has_Y02C, has_Y02D, has_Y02E, has_Y02P, has_Y02T, has_Y02W, has_Y04S, has_EvT_1_1, has_EvT_1_2, has_EvT_1_3, has_EvT_1_4, has_EvT_1_5*]

Under the main WIPO-digital category, we have 10 subcategories. Each patent family is accordingly tagged as the various combinations of the following 10 binary variables:

VarSet_WIPO = [*has_AI_ML, has_AutonmSys, has_BigData, has_CloudComp, has_IoT, has_Quantum, has_Robots, has_SmartCities, has_Blockchain, has_Virt_Aug_Reality*]

Under the main OECD JTag-digital, we have 13 subcategories. Each patent family is accordingly tagged as the various combinations of the following 13 binary variables:

VarSet_JTag= [*has_Jtag_1, has_Jtag_2, has_Jtag_3, has_Jtag_4, has_Jtag_5, has_Jtag_6, has_Jtag_7, has_Jtag_8, has_Jtag_9, has_Jtag_10, has_Jtag_11, has_Jtag_12, has_Jtag_13*]

The descriptions of the various technology subcategories referred to by the 37 mentioned above, are provided in an EXCEL file enclosed with the first dataset (see Zenodo Box).

Let us continue our demonstration with the French patents (regardless of year and NACE2 detail) focusing on the 10 subcategories under the WIPO-digital definition. As the first step, let us first see how many WIPO-digital patent families French inventors held from year 2000 onwards in all NACE2 industries

```
SELECT SUM(fr_nr_families) AS total_fr_nr_families
FROM Fractional_Counts_Detailed
WHERE Inv_Country = 'FR' AND has_WIPO_Digtl=1
```

The query yields the scalar value 5961.5633.

Now, let's see the number of families that were related to Artificial Intelligence and/or Machine Learning technologies:

```
SELECT SUM(fr_nr_families) AS total_fr_nr_families
FROM Fractional_Counts_Detailed
WHERE Inv_Country = 'FR' AND has_AI_ML= 1
```

which will provide us with the scalar 2839.153. If we repeated this last query, each time replacing the variable *has_AI_ML* with another one of the 10 binary WIPO-digital subcategory identifiers (in the set **VarSet_WIPO** as defined above), we would be able to construct (manually) the following table

WIPO subcategory	***total_fr_nr_families***
has_AI_ML	2839.153
has_AutonmSys	781.8278
has_BigData	106.8284
has_CloudComp	1027.042
has_IoT	164.0293
has_Quantum	163.9545
has_Robots	888.7085
has_SmartCities	122.1972
has_Blockchain	112.8826
has_Virt_Aug_Reality	168.0608
Column TOTAL	**6374.684**

Now we observe that the total (on the last row) is not the same as the total number of WIPO-digital patents held by French inventors, which, as we computed by the query before, was 5961.5633, which is substantially smaller than the column sum of the table above (i.e., 6374.684).

The obvious reason is the fact that in many cases, a patent family is tagged simultaneously in more than one sub-category under WIPO-digital (which is also the case within the other two main technology categories, EnvTech and JTag). Thus, the table we constructed above suffers from double-counting due to the multiplicity of the subcategory tags that happen at the patent level.

This kind of double counting would not be a concern when one needs patent family counts per subcategory (e.g., 2839.153 is the correct number of French patent families that are related to artificial intelligence and/or machine learning). Double counting only becomes a concern when one is interested in a mutually-exclusive breakdown over 'all' subcategories of a main category.

For those interested in such mutually-exclusive breakdowns into technological subcategories, we address this double counting issue by the introduction of a third layer of fractionalization (on top of inventor country and NACE2 industries) that takes into account the co-presence of the various technological subcategories. This introduces three more fractional count variables (one for each main technological category) into our new dataset. These new variables are alternatives for the variable `fr_nr_families` that we used so far in the example queries. In other words, the user should choose whether she wants to use `fr_nr_families`, which will imply double counts when J-tag, EnvTech or WIPO-digital are broken down into sub-categories, or use the new variables, which avoids such double counts.

The first new variable, `fr_nr_families_for_Green_BreakUp`, fractionalizes patents among the 14 OECD EnvTech subcategories (unconditionally on, i.e., disregarding, their sub-tagging in the other two main categories). For instance, if a patent family is tagged as related to the green subcategories `Y02A` (adaptation to climate change) and `EvT_1_3` (waste management), our new

variable fr_nr_families_for_Green_BreakUp will assign half of this patent family to Y02A and the other half to EvT_1_3. To put it more formally, for any given row i of the new dataset

$$\texttt{fr_nr_families_for_Green_BreakUp(i)} = \begin{cases} \texttt{fr_nr_families(i)} & \text{if } \texttt{has_Green(i)} = 0 \\ \texttt{fr_nr_families(i)}/N_G(\texttt{i}) & \text{if } \texttt{has_Green(i)} = 1 \end{cases}$$

where $N_G(\texttt{i}) = \textit{has_Y02A}(\texttt{i}) + \textit{has_Y02B}(\texttt{i}) + \textit{has_Y02C}(\texttt{i}) + \textit{has_Y02D}(\texttt{i}) + \textit{has_Y02E}(\texttt{i}) + \textit{has_Y02P}(\texttt{i}) + \textit{has_Y02T}(\texttt{i}) + \textit{has_Y02W}(\texttt{i}) + \textit{has_Y04S}(\texttt{i}) + \textit{has_EvT_1_1}(\texttt{i}) + \textit{has_EvT_1_2}(\texttt{i}) + \textit{has_EvT_1_3}(\texttt{i}) + \textit{has_EvT_1_4}(\texttt{i}) + \textit{has_EvT_1_5}(\texttt{i})$.

Following the same logic, our two other new variables fr_nr_families_for_WIPO_BreakUp and fr_nr_families_for_JTag_BreakUp fractionalize patent families respectively within the WIPO-digital and JTag main categories, once again exclusively (i.e., unconditionally on the breakdown within the other two main categories. Formally, for any given row i of the new dataset

$$\texttt{fr_nr_families_for_WIPO_BreakUp(i)} = \begin{cases} \texttt{fr_nr_families(i)} & \text{if } \texttt{has_WIPO_Digtl(i)} = 0 \\ \texttt{fr_nr_families(i)}/N_W(\texttt{i}) & \text{if } \texttt{has_WIPO_Dgtl(i)} = 1 \end{cases}$$

where $N_W(\texttt{i}) = \textit{has_AI_ML}(\texttt{i}) + \textit{has_AutonmSys}(\texttt{i}) + \textit{has_BigData}(\texttt{i}) + \textit{has_CloudComp}(\texttt{i}) + \textit{has_IoT}(\texttt{i}) + \textit{has_Quantum}(\texttt{i}) + \textit{has_Robots}(\texttt{i}) + \textit{has_SmartCities}(\texttt{i}) + \textit{has_Blockchain}(\texttt{i}) + \textit{has_Virt_Aug_Reality}(\texttt{i})$.

and also

$$\texttt{fr_nr_families_for_JTag_BreakUp(i)} = \begin{cases} \texttt{fr_nr_families(i)} & \text{if } \texttt{has_JTag_Digtl(i)} = 0 \\ \texttt{fr_nr_families(i)}/N_W(\texttt{i}) & \text{if } \texttt{has_JTag_Dgtl(i)} = 1 \end{cases}$$

where $N_J(\texttt{i}) = \textit{has_Jtag_1}(\texttt{i}) + \textit{has_Jtag_2}(\texttt{i}) + \textit{has_Jtag_3}(\texttt{i}) + \textit{has_Jtag_4}(\texttt{i}) + \textit{has_Jtag_5}(\texttt{i}) + \textit{has_Jtag_6}(\texttt{i}) + \textit{has_Jtag_7}(\texttt{i}) + \textit{has_Jtag_8}(\texttt{i}) + \textit{has_Jtag_9}(\texttt{i}) + \textit{has_Jtag_10}(\texttt{i}) + \textit{has_Jtag_11}(\texttt{i}) + \textit{has_Jtag_12}(\texttt{i}) + \textit{has_Jtag_13}(\texttt{i})$.

With these three new variables in place, let us revisit our last example of the breakdown of the WIPO-digital patents by French inventors into the WIPO subcategories, so that we can see how one avoids double counting when mutual-exclusivity is a concern. All we need to do is to replace the variable name fr_nr_families with fr_nr_families_for_WIPO_BreakUp:

```
SELECT SUM(fr_nr_families_for_WIPO_BreakUp) AS
total_fr_nr_families_for_WIPO_BreakUp
FROM Fractional_Counts_Detailed
WHERE Inv_Country = 'FR' AND has_AI_ML= 1
```

which will provide us with the scalar 2684.0186. Due to the new fractionalization layer (avoidance of double counting), this number is lower than in the above query with fr_nr_families. Just as before, if we repeated this query, each time replacing the variable has_AI_ML with another one of the 10 binary WIPO-digital subcategory identifiers, we would be able to construct the same table as before. We juxtapose this new table with the old one:

WIPO subcategory	total_fr_nr_families	total_fr_nr_families_for_WIPO_BreakUp
has_AI_ML	2839.153	2684.0186
has_AutonmSys	781.8278	700.1144
has_BigData	106.8284	76.5284
has_CloudComp	1027.042	985.6406
has_IoT	164.0293	147.6117
has_Quantum	163.9545	156.3323
has_Robots	888.7085	832.7623
has_SmartCities	122.1972	120.067
has_Blockchain	112.8826	102.2946
has_Virt_Aug_Reality	168.0608	156.1934
TOTAL	**6374.6840**	**5961.5633**

Observe that, unlike the case before, the new fractional figures (in the third column) add up perfectly to the total number of WIPO-digital patent families held by French inventors, which we had computed previously also as 5961.5633.

Let us round up this section by re-emphasizing that the three new variables introduced in this version of the dataset are created in view of one specific purpose: to create a mutually-exclusive breakdown of patents simultaneously over 'all' subcategories of a main technological classification. For any purposes other than that, the 'old' variable `fr_nr_families` should remain as the basis of aggregation. To give an example: the actual number of French patents related to AI/ML is computed on the basis of the variable `fr_nr_families`, and is equal to 2839.153 as indicated in the second column of the table above. The corresponding figure 2684.0186 on the second column which is inevitably smaller.

4. Technology flow matrices

The technology flow matrices are based on the same system of fractional counting and tagging as explained so far. The main principle for identifying flows is that of patent citations: if there is a citation by any member of patent family B to any member of patent family A, we identify a technology flow from A to B, which will be recorded fractionally to the countries and sectors assigned to the families A and B. We will use only non-'quality-filtered' patent families (see Section 1) to construct the knowledge flows dataset.

An individual patent that is either the origin or destination in a citation pair is assigned to one or more countries, and one or more NACE2 sectors. It can also be tagged, or not, in J-tag, in WIPO-digital, and in EnvTech (green). In the current version of the technology flow matrices, the latter classifications (J-tag, WIPO-digital and EnvTech) are binary, so each patent family has three binary variables: J-tag-digital, WIPO-digital, and EnvTech where a 1 indicates J-tag digital, WIPO-digital and EnvTech-green, respectively. This means that we do not currently include any information about the sub-categories of J-tag, WIPO-digital or EnvTech in the technology flow matrices.[1] At the level of a (cited or citing) patent family or application, all this is similar to the method used for fractional counting in the first dataset for patent counts (Section 2).

[1] This information can be included in the future, but including a large number of sub-categories may be problematic because the dimensionality of the resulting matrices increases rapidly with the number of sub-categories.

As in the case of patent counts, our procedure starts by enumerating the 'varieties' of a patent application in terms of its sectors and inventor countries, using the fractional weights for NACE2 sectors and inventor location countries. As before, this is aggregated to the patent family level by adding up and fractionalizing over the members (applications) of a family. Next, we construct a set of citation pairs, at the patent family level. These are pairs of cited and citing patent family that satisfy our timing criteria, i.e., the cited patent has earliest application date in 2000 or later, and the citing or cited patent falls within a fixed citation window.

The citation window applies to the time distance between cited and citing patents, and we apply this in two different ways, leading to two versions of the technology flows dataset. In both cases, we use a citation window of 6 years. In the first way of looking at the citation window, which can be seen as applying a forward-looking perspective, for a cited patent in year T, only citations in the period from T until T+5 (both inclusive) will be counted.[2] In the backward-looking perspective, for a citing patent from year T, we only consider cited patents from T–5 to T (again, inclusive).

We use the fixed citation window to prevent earlier patents from accumulating more citations, and hence 'generate' more flows, than later patents. Whether the backward- of forward-looking dataset is used, depends on the context in which the data will be used, and is for the user to decide. Note that in the forward-looking perspective, due to the truncation of the dataset in 2023, the actual citation window will be shorter than 6 years for patent families in 2018 and later. Similarly, in the backward-looking perspective, truncation occurs in 2000, which implies that the citation window in the first cohort is limited (e.g., citing patents in 2000 only cite patents in the same year).

The notion of a citation pair is central to the procedure of building the technology flows data. We use citations between families, and identify such a citation if at least one member of a family cites at least one member of another family. The following table shows an example of such a citation pair for patent families A (cited) and B (citing), with their fractional weights listed.

	Family A (cited)	**Family B (citing)**
Weight Inventor country BE	0	0.2
Weight Inventor country NL	0.15	0
Weight Inventor country FR	0.85	0.8
Weight NACE2 26.1	0.4	0.25
Weight NACE2 26.2	0.6	0.75
J-tag (binary)	1	1
WIPO-digital (binary)	1	0
EnvTech green (binary)	1	1

These patents are (partially) invented in Belgium, France and the Netherlands, and in NACE2 sectors 26.1 and 26.2. Each family has 2 inventor countries and 2 sectors, hence there are 4 country – sector combinations for each family. A key insight is that this makes $4 \times 4 = 16$ citation combinations in terms of countries and sectors. These 16 combinations are enumerated in the following table, using colour-coding for the 4 combinations at the cited and citing level:

[2] It is T+5 for a 6-year window because also citations in T itself are considered. Although, theoretically, there should be no cases where the cited patent has a later date (year) than the citing patent, this does happen in PatStat. We exclude those cases.

Origin (cited) (country-NACE2)	Weight for origin	Destination (citing) (country-NACE2)	Weight for destination	Weight for citation
NL-26.1	$0.15 \times 0.4 = 0.06$	BE-26.1	$0.2 \times 0.25 = 0.05$	$0.06 \times 0.05 = 0.003$
NL-26.1	$0.15 \times 0.4 = 0.06$	BE-26.2	$0.2 \times 0.75 = 0.15$	$0.06 \times 0.15 = 0.009$
NL-26.1	$0.15 \times 0.4 = 0.06$	FR-26.1	$0.8 \times 0.25 = 0.2$	$0.06 \times 0.2 = 0.012$
NL-26.1	$0.15 \times 0.4 = 0.06$	FR-26.2	$0.8 \times 0.75 = 0.6$	$0.06 \times 0.6 = 0.036$
NL-26.2	$0.15 \times 0.6 = 0.09$	BE-26.1	$0.2 \times 0.25 = 0.05$	$0.09 \times 0.05 = 0.0045$
NL-26.2	$0.15 \times 0.6 = 0.09$	BE-26.2	$0.2 \times 0.75 = 0.15$	$0.09 \times 0.15 = 0.0135$
NL-26.2	$0.15 \times 0.6 = 0.09$	FR-26.1	$0.8 \times 0.25 = 0.2$	$0.09 \times 0.2 = 0.018$
NL-26.2	$0.15 \times 0.6 = 0.09$	FR-26.2	$0.8 \times 0.75 = 0.6$	$0.09 \times 0.6 = 0.054$
FR-26.1	$0.85 \times 0.4 = 0.34$	BE-26.1	$0.2 \times 0.25 = 0.05$	$0.34 \times 0.05 = 0.017$
FR-26.1	$0.85 \times 0.4 = 0.34$	BE-26.2	$0.2 \times 0.75 = 0.15$	$0.34 \times 0.15 = 0.051$
FR-26.1	$0.85 \times 0.4 = 0.34$	FR-26.1	$0.8 \times 0.25 = 0.2$	$0.34 \times 0.2 = 0.068$
FR-26.1	$0.85 \times 0.4 = 0.34$	FR-26.2	$0.8 \times 0.75 = 0.6$	$0.34 \times 0.6 = 0.204$
FR-26.2	$0.85 \times 0.6 = 0.51$	BE-26.1	$0.2 \times 0.25 = 0.05$	$0.51 \times 0.05 = 0.0255$
FR-26.2	$0.85 \times 0.6 = 0.51$	BE-26.2	$0.2 \times 0.75 = 0.15$	$0.51 \times 0.15 = 0.0765$
FR-26.2	$0.85 \times 0.6 = 0.51$	FR-26.1	$0.8 \times 0.25 = 0.2$	$0.51 \times 0.2 = 0.102$
FR-26.2	$0.85 \times 0.6 = 0.51$	FR-26.2	$0.8 \times 0.75 = 0.6$	$0.51 \times 0.6 = 0.306$

As the table shows, we simply multiply the weight for the origin (cited patent family) and destination (citing patent family) to obtain the weight for the citation. In this way, the citation weights will add up to exactly 1 over the 16 possible combinations. Also, if we sum the citation weights over each variety of cited or citing family, we find the origin and destination weights.

This brings us to the question how the ultimate technology flow dataset is built. This is done by first building a table like the one above for all citation pairs, and then summing over citation weights in all those individual-citation-pair tables together. In doing this, we use also all possible combinations of the binary J-tag, WIPO-digital and EnvTech-green variables. This adds $2^3 \times 2^3 = 64$ possible combinations for the 3 tags.[3] Note that in case of the above table, only one of those combinations applies (1-1-1 with 1-0-1, for J-tag-WIPO-digital-EnvTech). By adding over more patent citation pairs, a larger number of those 64 possible combinations (possibly all 64) will appear.

We also add up citation weights over time cohorts. These are 6-year periods as in 2000-2005, 2006-2011, 2012-2017 and 2018-2023, where the years refer to the cited patent family in case of the forward-looking perspective, and the citing patent family in case of the backward-looking perspective. The fixed citation window is applied here. This means that in the forward-looking perspective, in the first cohort the earliest citing patent is from 2000 (citing a patent from the same year), and the latest citing patent is from 2010 (citing a patent from 2005). This holds in a similar fashion for the later cohorts. For the backward-looking perspective, the earliest citing patent is from 2000.

It is important to realize that this procedure implies that patent families that are cited more often also add more to the total sum of citation weights, simply because they appear in more citation pairs, and each citation pairs contributes a value of exactly 1. Of course, this also holds for citing patent families: those that cite more will contribute more in the sum of citation weights. This represents the idea that each citation can be seen as a unit-flow of knowledge.

[3] Here we can see what a large number of sub-categories for J-tag, WIPO-digital and Envtech implies. With 13 J-tag categories, 10 WIPO-digital categories and 14 EnvTech categories as in the second patent counts dataset above, we have $13 \times 10 \times 14 = 1820$ categories for both citing and cited, which makes $1820^2 = 3.3124\ million$ combinations, which comes on top of the combinations of country and NACE2 for both citing and cited.

Table 1. Example rows of the technology flows dataset

Forward-looking perspective

time_cohort_cited_fam	Cited_Inv_Country	Cited_nace2_code	Cited_has_Green	Cited_has_JTag	Cited_has_WIPO_Digtl	Citing_Inv_Country	Citing_nace2_code	Citing_has_Green	Citing_has_JTag	Citing_has_WIPO_Digtl	fr_nr_citations
2000_2005	JP	18.1	0	1	0	KR	31	0	0	0	0.1142290
2006_2011	US	10	1	0	0	US	28.23	1	0	0	0.2539516
2012_2017	IL	28.25	1	0	0	CN	26.1	0	0	0	0.0515873
2018_2023	RU	28.23	0	0	1	IN	26.3	0	1	0	0.0055555

Backward-looking perspective

time_cohort_citing_fam	Cited_Inv_Country	Cited_nace2_code	Cited_has_Green	Cited_has_JTag	Cited_has_WIPO_Digtl	Citing_Inv_Country	Citing_nace2_code	Citing_has_Green	Citing_has_JTag	Citing_has_WIPO_Digtl	fr_nr_citations
2000_2005	ES	26.51	0	1	0	US	28.95	1	1	1	0.0009259
2012_2017	NO	30	0	0	0	GB	27.3	0	1	0	0.0685185
2006_2011	CA	27.3	1	1	0	DE	26.1	1	1	0	0.0024337
2018_2023	CZ	62	0	1	0	JP	29.1	0	1	0	0.0018994

The format of the ultimate table is illustrated in Table 1. This contains just 4 example rows, the actual table has 33,128,317 rows in case of the forward-looking perspective, and 34,745,424 rows in case of the backward-looking perspective. Each row has a combination of country and NACE2 sector for cited and citing patent families, a combination of the binary variables for J-tag, WIPO-digital and EnvTech (green), also for cited and citing patent, a column indication the cohort, and a column containing the citation weight (`fr_nr_citations`).

The following example query extracts an aggregated technology flows dataset that only includes green-to-green knowledge flows for cohort 1, from the forward-looking dataset (time windows for the citing patents):

```
SELECT
    Cited_Inv_Country,
    Cited_nace2_code,
    Citing_Inv_Country,
    Citing_nace2_code,
    SUM(fr_nr_citations) AS total_fr_nr_citations
FROM
    Fractional_Citations_6_yearlyCohorts_Of_CitedFam
WHERE
    Cited_has_Green = 1
    AND Citing_has_green = 1
    AND time_cohort_cited_fam = '2000-2005'
GROUP BY
    Cited_Inv_Country,
    Cited_nace2_code,
    Citing_Inv_Country,
    Citing_nace2_code
```

The `Cited_has_Green`, `Cited_has_Jtag`, `Cited_has_WIPO_Digtl` and the citing counterparts `Citing_has_Green`, `Citing_has_Jtag`, and `Citing_has_WIPO_Digtl` can be used to extract other J-tag, WIPO-digital and EnvTech combinations. For example, this query will create an aggregated green-to-WIPO-digital matrix from the same dataset:

```
SELECT
    Cited_Inv_Country,
    Cited_nace2_code,
    Citing_Inv_Country,
    Citing_nace2_code,
    SUM(fr_nr_citations) AS total_fr_nr_citations
FROM
    Fractional_Citations_6_yearlyCohorts_Of_CitedFam
WHERE
    Cited_has_Green = 1
    AND Citing_has_WIPO_Dgtl = 1
    AND time_cohort_cited_fam = '2000-2005'
GROUP BY
    Cited_Inv_Country,
    Cited_nace2_code,
    Citing_Inv_Country,
    Citing_nace2_code
```

As a final example, this query will create a technology flows dataset that includes only knowledge flows from green-but-not-digital to WIPO-digital-to-not-green patent families:

```
SELECT
    Cited_Inv_Country,
    Cited_nace2_code,
    Citing_Inv_Country,
    Citing_nace2_code,
    SUM(fr_nr_citations) AS total_fr_nr_citations
FROM
    Fractional_Citations_6_yearlyCohorts_Of_CitedFam
WHERE
    Cited_has_Green = 1
    AND Cited_has_WIPO_Dgtl = 0
    AND Citing_has_WIPO_Dgtl = 1
    AND Citing_has_Green = 0
    AND time_cohort_cited_fam = '2000-2005'
GROUP BY
    Cited_Inv_Country,
    Cited_nace2_code,
    Citing_Inv_Country,
    Citing_nace2_code
```

Finally, we want to point to the option of normalizing the raw numbers on technology flows. For example, suppose that we found, by one of the queries above, that in cohort 1, the knowledge flow from patents invented in France in NACE2 = 26.1 to patents invented in Belgium in NACE2 = 26.2 is equal to 100. The user may (or not) want to normalize this flow, for example by the total number of patent families invented in France in NACE2 = 26.1, in order to reflect the knowledge flow per-origin (cited) units. To this, end the user would have to find the number of patent families invented in France in NACE2 = 26.1 during the cohort 1 period in the first patent count dataset as explained above. If that number of patent families would be equal to 1,000, then the normalized knowledge flow would be 0.1.

Or, the user may want to normalize by the number of destination patent families, in this case patent families invented in Belgium in NACE2 = 26.2. Again, the user would have to find the number of patent families invented in Belgium in NACE2 = 26.2 during the cohort 1 period in the first patent count dataset as explained above. Either type of normalization may or may not be necessary, depending on the specific context in which the knowledge flows dataset is used.