FURTHER DEVELOPED GCTS RESEARCH CONCEPT
Készült az EFOP-3.4.5-VEKOP-17-2017-00001 azonosítószámú, Rendszerszintű fejlesztések és hozzáférés bővítését szolgáló ágazati programok a felsőoktatásban kiemelt projekt DPR Pillér megvalósítása során.

*Szerző: Eszter Ádám*

Felelős kiadó:
Dr. Gloviczi Zoltán

© Oktatási Hivatal, 2020

Az Oktatási Hivatal kiadványa az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg. A kiadvány adatainak bármilyen formában történő nyilvános közötteletéhez vagy felhasználásához a kiadó előzetes írásbeli hozzájárulása szükséges.
# TABLE OF CONTENTS

Introduction ............................................................................................................................ 2

1. Graduate career tracking system (GCTS) ........................................................................... 5

  Tracking drop-out and graduate students ................................................................. 5
  Tracking early labour market life paths ................................................................. 8
  Career tracking analyses ....................................................................................... 9

Research modules of Graduate Career Tracking System .................................................. 13

Integration of Administrative Databases (IAD) ................................................................. 13

  Conceptual model ..................................................................................................... 14

  A GCTS IAD core variables as a result of renewal by information provider partners .... 15

Method for data checking ............................................................................................. 18

  Data process methodology ....................................................................................... 18

  Data visualisation renewal ....................................................................................... 18

Fresh graduate research ................................................................................................. 19

  Contents and structural renewal of the core questionnaire ....................................... 21

System elements proving the needs for the register-based survey .................................. 25

Higher Education career tracking .................................................................................. 28

New research methodology ......................................................................................... 29

  Professional and technical dimensions of methodology: ......................................... 29

  Professional dimension ............................................................................................ 29

Research findings – career tracking of IT educational sector students .......................... 32

Dissemination .................................................................................................................. 33
Introduction
Graduate career guidance has a ten year history in Hungary. It was founded in 2008 by Educatio Non-profit Ltd. (Educatio Társadalmi Szolgáltató Nonprofit Kft), and the first institutional data were collected in 2009. Over the last decade, career guidance environment and the external and internal circumstances have also changed significantly.

In Hungary, only the Educational Authority does systematic graduate career tracking yearly, based on data integration and including all graduates (in accordance with all applicable laws and regulations, and within the frames of very strict legislation).

Within the frames of the Human Capacities Development Programme (HCDOP) 3.4.5 Higher Education Sectoral Programme Aiming at System Level Development and Widening Accessibility project, graduate career tracking system has gone / is going through fundamental changes. The functioning of graduate career tracking system is enabled by Act CCIV of 2011 on National Higher Education and 87/2015 (IV.9) Government Decree about its implementation, and Act CI of 2007 on ensuring access to data for decision making and 335/2007. (XII.13.) Government Decree about its implementation.

The sub-researches of graduate career tracking have improved immanently, however, separately over the last decade, which resulted in overlaps of bordering fields in different research phases – meanwhile research profile clean-up and precise specification of conceptual models just did not happen. Setting clear frames and borders was an important challenge in the research development, parallel with finding possible interfaces and synergies, and developing relevant methodology for them. The main goal of the project was to renew the career tracking system structure and the contents of its modules: we aimed to set a systemic research structure: we eliminated duplications in the research modules, merged data integration and survey methods into a new form of research through the register based survey pilot.

Changing was inevitable. One of the main problems of the current system is that parallel with the growing data integration research modules, the Fresh graduate research ‘research position’ has weakened, and what was its systemic advantage at the time of introduction has become its disadvantage: it provides solutions to several questions, however, each one is rather superficial (due to limited data collection possibilities). Due to data gaps, there were no data available about e.g. how much a fresh graduate earns, and how long they spend with job-seeking. Through the strengthening of data integration, these factual data can be examined more thoroughly in register based researches, data reporting is faster, more exact and reliable from these databases than from those questionnaires based
on self-declaration having less items. This shift in emphasis did not bring automatically the renewal of the Fresh graduate questionnaire in terms of enabling (in a questionnaire form) real deep analysis of special topics by clearing off duplications, and such data can be analysed better in registers. The early value of the Fresh graduate research, namely that it provided (superficial) information of everything has become its disadvantage.

When developing the career tracking model there was only one tool available: online questionnaire. The questionnaire has tried to serve career tracking needs best: consequently, it touched upon all relevant topics in short, resulting in superficial and long-running references in topics. In 2011, a new pilot research module was introduced: data integration based career tracking. Methodologically, it is a very important difference that data integration is based on factual data stored in administrative databases, while the online questionnaire is on self-declaration responses. The development of data integration module brought the needs for delicately differentiating between datasets that could be subject to a research: databased researches provide better and more precise reflection to information about labour market status and income; in the meantime, an online questionnaire is suitable e.g. for assessing employment perceptions, labour market expectations, efficiency of education and training in the labour market.

The overall renewal of the data integration research module was implemented in 2018 within a subproject, followed by the renewal of Fresh graduate research after the necessary diversification of research profiles in 2019, and then by 2020 will come the refreshed Student research as well.

In the Human Capacities Development Programme (HCDOP) 3.4.5 Higher Education Sectoral Programme Aiming at System Level Development and Widening Accessibility project, during the renewal of graduate career tracking system we have developed two innovative research modules: the Higher Education Career tracking module that provides the possibility for examining the input side and the register based survey that is capable of combining the advantages of data integration and online researches.

Higher Education Career tracking is different from the GCTS IAD research in its baseline. The population of the Graduate Career Tracking System – Integration of Administrative Databases (GCTS IAD) research module is the graduates of a given year. The Higher Education Career tracking examines parallel the education / training and labour market life paths of first year students, thus we gain more information about working while studying and its impact, and the connections between student and working life paths.
We implemented the register based survey research as a pilot project in order to renew the Fresh graduate research structurally. Online questionnaire queries are limited in terms that the time range given for filling out cannot be overloaded with requiring base information, however, if the number of background variables is limited, assessment will be of limited possibilities as well. The register based survey provides a solution for this: we link administrative data stored in the Higher Education Information System (HEIS) with the responses given in the questionnaires in an anonymised way. The research managed by the Educational Authority has proved exceptionally good results in response rate (nearly double in comparison to that of the current solution applied by higher education institutions; without a backup campaign we achieved a 23% response rate).

The renewed Hungarian graduate career tracking system supplemented by research models is so complex that it is unique to not only Europe but worldwide, too.

The complex system requires three fundamentals for its managing:
- suitable legislation
- high quality administrative databases and
- special professional knowledge, knowhow.

These three conditions are available in Hungary, which is unique even in Europe, enabling the operation of a complex, four-module research system, which contains research modules requiring big data type, data mining skills and questionnaire researches based on ‘classic’ sociological knowledge.

As a result of Human Capacities Development Programme (HCDOP) 3.4.5 Higher Education Sectoral Programme Aiming at System Level Development and Widening Accessibility traditional career tracking modules renewed, a completely new data processing knowhow has been introduced, new research modules have been developed, central career tracking questionnaires renewed in their contents.
1. Graduate career tracking system (GCTS)

A graduate career tracking system aims to follow up and analyse higher education drop-outs’ and graduates’ early labour market life paths. Hungarian career tracking is a uniquely complex system consisting of more research modules, whose special value is the synergy of the questionnaires and data integrations evaluations.

Educational Authority operates a complex career tracking system (consisting of 4 research modules), thus the GCTS is basically a set of researches built upon diverse conceptual models implemented annually through a diverse population and different data collection methods. Some parts of the research module are based on factual data (via connecting administrative databases), other parts are based on students’ self-declaration, online questionnaires results (which are suitable for studying motivations, attitudes).

The operation of the graduate career tracking system is regulated and supported by different legislation. There is no singular overall law that regulates career tracking system in general: the legal basis of some parts of the research modules are stipulated by the national law on higher education and Articles of certain regulations, other parts of the research modules are regulated by other laws. Educational Authority, particularly the Higher Education Analysis Department is responsible for the maintenance and managing of the career tracking system (in other words the implementation of the researches), data reporting and analysis.

The graduate career tracking system aims to follow up and analyse higher education drop-outs and graduates early labour market life paths.

Tracking drop-out and graduate students

Determining the notion of “completion of higher education studies” is not as simple as it seems at first sight. Who is a graduate? Those who have completed all state exams? Degree holders? Those who have completed all state exams, but do not hold a degree yet, however completed all major exams? Depending on what answers we give to these questions, we end up with diverse results about the number of graduates. Those who finished their studies, have they completed them? How do we define those who do not get as far as the state exams, or without the compulsory language exam will not be awarded a degree – in other words fail to complete their higher education studies? What to name them, which category shall we include them? These are important questions to a researcher, which need a two-stage approach to answer: on the one hand, we need to define precisely the working definition, on
the other hand, once we can specify the target group on the basis of the definition, we need to identify those sets of data in the databases that describe the target group precisely.

We identify the target group in the career tracking system by requesting that population from the Higher Education Information System (that contains all data of every higher education student) which can be regarded as graduates of a given year. Since we use Higher Education Information System data for our analysis, we aim to tailor our terminology in accordance with HEIS terminology. In most cases we manage to do so, however, at this stage we need to highlight upon a researcher’s dilemma. The Higher Education Information System is an administrative database, designed for storing data of administrative procedures in the most suitable way. Administrative and research database logics are sometimes different, consequently it leads to difference in giving definitions as well.

As researchers, we are interested in the following: HEIS simply tells us who graduated in 2018. However, HEIS, as it is an administrative database, cannot provide an answer. Since HEIS is an administrative database, it cannot understand this request; it does not need information of graduates this way and in this form. Higher Education Information System cannot interpret the notion of “graduation”. It understands high performing students, drop-outs, and those finishing studies for several other reasons – and stores these information in one field (we might as well call these: variables). In another field it stores information of the degree, but not in the form we researchers would need (namely binary variable categories: yes/no), but by dates, which refer to the date of issuing the degree.

In order to understand HEIS logic, we need to understand the needs and principles of the sector which were taken into consideration when this database was set up. If HEIS had been established as a research database, it may store answers differently. But HEIS is an administrative database; its set of data has specific aims – which are not for research purposes in the first place. Researchers need to learn how to find answers to their hypothesis, which variables they need to analyse simultaneously in order to find the answer, how to modify variables to get particular answers to particular questions.

Research of administrative data always starts with getting to know the given database which provides data. For example, in the National Tax and Customs Administration database we can find negative income. Researches need to be able to differentiate whether it is data error or not – if the latter is the case, we must be able to interpret the data in the field!
In this case the notion of “completion” is different in the researchers’ and data managers’ understanding. In case of successful graduates successful completion does not necessarily mean that they will be awarded a degree. Successful completion means the student has taken provisional certificate of graduation (abszolútium) exams at least; moreover, completed outcome exams (final exam) as well.

The difference, and the nature and reason for different data storage arise from the fact that successful completion of studies does not automatically lead to obtaining a degree – since the degree is awarded only after the completion of all final (state) exams and particular number and level of language exams. Thus what is an ordinal scale for us, researchers – will be two separate fields with own data storing logic in the administrative databases:

- **HKPZ_MSZ_TIPUS_NEV** Reasons for terminating training
- **HKPZ_OKLEVEL_DAT** Date of degree obtained

**Interpretation of HEIS fields**

*Reason for terminating studies:*

Successful completion, which shows how a student completed their higher education studies, is among reasons for terminating training in HEIS. The two main groups of reasons for termination are:

**Successful completers**
- Successful outcome exam (KTB)
- Completion of studies, obtaining a leaving certificate without outcome exam (KVB)

**Drop-outs**
- Exclusion after disciplinary actions (KKI)
- Disagree to pay tuition fee in case of transfer (KKN)
- Fail to register more times than allowed (KMB)
- Arrears in education (KMF)
- Self-notification about terminating training (KSK)
- Fail to fulfil educational requirements (KTO)
- Exceed the number of allowed exam failures and exam repeats (KVS)

**Others**
- Transfer to another institution on student’s request (KAK)
- Transfer to another institution on closure (KAM)
o Health unfitness (KFE)
o Death (KHA)
o Termination due to institution closure (KIM)
o Transfer to successor institution for succession (KJU)
o Transfer to different department or faculty within the same institution (KMK)
o Completion of a study period (KRT)
o Completion of preparatory courses (KEL)
o Completion of refresher training (KRI)
o Transfer to another institution abroad on student’s request (KKU)
o Continuing studies.

*Obtaining a degree*

Having obtained/lacking of a degree is in a singular data field in HEIS. We regard a graduate as one with a degree if the particular field is filled out with the date of obtaining.

Thus we know in case of every single student what reason they have for terminating studies in a given year, and we also know if they are awarded a degree in a given year, or any other time (by referring to the date).

In order to define our target group – in other words who completed studies in a given year – we collect Successful completers in the HEIS dictionary. As we also analyse drop-outs, we define target groups on the basis of further termination reasons: e.g. persons of all subclasses within the drop-outs category, but persons of other category are of no condition to the query. It is a special task to identify those within the target group who have obtained a degree – having said that the date of obtaining field within the HEIS. If we want to identify those having obtained a degree in our target group, we look for whether the date field is filled out or not. If the field is filled out – then those are graduates.

*Tracking early labour market life paths*

There is no existing standard for defining the notion of early labour market life path. At the beginning, career tracking analysis examined labour market status typically in a 1-3-5-year perspective, while today – specially in data integration modules – we may come across longer, even a ten-year perspective tracking or over. Questionnaire queries are limited in terms that 5-10 years after graduation email
databases are not reliable anymore (except if due to the alumni network there is a continuous contact with former graduates). However, the reliability of the institution databases depends on the database itself: there are better and less maintained databases, which may well serve the research needs of a given institutions but for a sectoral planning their good enough quality is not enough. Thus in Hungarian career tracking system, in case of the questionnaire query we do only a maximum of 5-year tracking, and from 2018, in case of data integration modules we increased the length of tracking for 7 years – but in accordance with international standards we may do tracking in a longer perspective.

It is an exciting question what is the date when the tracking itself is interesting or worth being done. Among international queries we find such ones that function as online questionnaire queries 7-9 months after graduation. In the Hungarian system, data collection usually takes place in May-June in the commencing year after the completion of the final study year, thus graduates receive the first questionnaire 10-12 months after completing their last semester (all in all their studies). The online questionnaire is sent out by the higher education institutions to their own graduates, and the time of sending out depends on the given institution (we know about such ones that consult graduate data base in the autumns).

Until 2018, we approached graduates through institutional mediation for filling out the questionnaire 1, 3, 5 years after the completion of studies. In the questionnaire of three years after studies the answers and trends did not represent a high quality of added value, thus from 2019, in compliance with international standards we approach alumni one and five years after the completion of studies, quitting the three-year query in the career tracking system.

Career tracking analyses

Prior to analysing career tracking data, we need to process and systemise data to make them suitable for research analysis. Different researches require diverse data processing and data storing skills and operations. The data of the Fresh graduate research and the Student research can be stored in Excel spreadsheets and SPSS, can be easily analysed (even in case of databases merged years retrospectively). In the online questionnaire research an average of 20.000 respondents are involved annually with a 10-12% of a response rate. In the data integration process the population is far greater; in 2017 we queried data of 140.000 persons, and due to the longer tracking time and the expansion of the target group we stored data about 604.000 persons in the database in 2018. The data amount of the latest data integration process
was over 2.5 billion single records – which neither Excel nor SPSS can manage – consequently we need other methodological procedures for processing, cleaning, and for the analysis itself (for these reasons we call data integration databases ‘small big data’ among us).

The new millennium has introduced immeasurable expansion of data collection possibilities. Big data, a new ideal, gains more grounds as a new research paradigm in market research as well as in scientific research, thus in social sciences, too. Big data is a novel way to data collection and data process, which is brought about in social sciences because – and it is no exaggeration to say – every single aspect or happening of our life produces some kind of data. Data set providing factual information of our life can comprise either personal contribution (e.g. when we enable tracking function on our smart phone and allow a third party to use all data generated this way), or a set of information from administrative databases, registers where stored in compliance with relevant legislation. In our life, several data in several places are generated about us as citizens. When we start in a new workplace, our employer reports to different authorities where we work and in what position. Year by year, we prepare a tax return listing all our incomes. When we move abroad or start a job abroad, we must report this fact, too, naming which country we will move to. The generated data set “contains information about people’s behavior”, in comparison to the “traditional way” of collecting social research data, which is based on observation, self-declarations. (Interview with Alex Pentland by edge.org, 8 August 2012, https://www.edge.org/conversation/alex_sandy_pentland-reinventing-society-in-the-wake-of-big-data.) Significant features of big data are the huge amount of data, their extremely fast procession and a great variety. In other words of the specific big data research language, these are the 3Vs, namely volume, velocity and variety. Besides them, these days we may come across with the fourth V, namely veracity; moreover, in some articles even with the fifth one called value.

The nature of data collection changes the underlying logic and methodology of process and analysis, widening the possibilities of data usage with new areas, e.g. with prediction. It is an exciting issue whether we can talk about social sciences big data bases separately within big data, or not. Administrative registers, which may serve as a basis for social researches big data, are also 21st century products. Their existence and widespreadness are far more dependent on culture than scientific data collection, or databases generated within global networks. In terms of their existence, governmental intention for their setting up (and the possibility for the integration of different databases), a decent data culture for data collection, the IT background for data collection and storage, and specific professional skills for data process are all vital. These factors must be available in order to set up such administrative registers that serve as a possible basis for the social research big data type of researches. The state, governmental intention is one of the most crucial ones. In the meantime, in more European
countries, besides the existing governmental support, data culture is in such conditions that it is rather pointless to have high expectations of administrative registers. Therefore, such a database is generated, if generated at all, this way that is useless for both administrative and research purposes.

The social sciences big data databases generated from administrative data are not based on direct questioning of persons, but on anonymously connected information stored about persons in administrative registers. Consequently, the important difference is that they contain factual data instead of personal opinions, thus we may set higher expectation towards them in terms of veracity, data hygiene, clarity and identifiability of data, continuous maintenance and updating, or a low level of data gaps.

The IT conditions and the underlying logic of data storage are not different in the case of big data and the huge databases of social sciences researches, since in both cases we are talking about a massive amount of data. However, it is more than challenging to define precisely what we can call as a great volume. Are data management and data process different in case of e.g. a weather forecast data, where hundreds of data is generated every second, and billions of specific data in a social research database? Regardless if the answer is yes or no, it is a fact that the common traditional database managing or data analysis software programmes cannot possibly process this amount of data. Far deeper IT and database managing skills are necessary.

Velocity is also an important common denominator – it is important in scientific and business researches mainly because newer and newer data return continuously into the database. In social researches, it is important in terms that in there instant data processing resulting in almost real time statements is a must. Variety also challenges business and social researches big data database analysts. Complex data, generated in different fields need to be structured and then connected to each other; data of different fields need to be integrated with each other, in other words, data, regardless of their sources, need to be arranged into a coherent database that is suitable for analytical purposes.

Integrated databases created this way make it possible to open up such correlations that single databases could not show separately (e.g. how the major of a given institution a student graduated at has an impact later on their personal monthly incomes). “Small big data” (SBD), which social research works with, plays an important role in the big data research wave; SBD as based on the integration of administrative databases means an especially exciting field of usage and processing state data assets, while also bringing society on board. In our terms, we consider those databases a big data-type in which the number of examinable cases is in the millions – this is the amount of data where the spreadsheets and database managing, statistical data analysis softwares widespread in social scientific survey-type of researches are of no use anymore. Obviously, the following IT skills are also required throughout the
whole research process: data warehouse and the big data approach, competence of data quality control and profiling, application of relational and data market-modelling tools, understanding of data integration methods between systems, handling of extract, transform, load (ETL) procedures, application of interactive business analysis tools as well as data visualisation and dashboarding skills.

Social sciences big data exist in “small big data” form in Hungary too, where unique legislation and data culture conditions support their existence, and guarantee a continuous development enabling one of the best graduate career tracking systems of Europe to operate in Hungary. The higher education database stored by Higher Education Information System (HEIS) is absolutely unique in Europe, and probably in the world as well. The research programme of Graduate Career Tracking System – Integration of Administrative Databases – GCTS IAD operated by the Educational Authority integrates data from six administrative register databases each structured by completely different logic, while also enables them to extract valuable information instantly. The data integration modules of Graduate Career Tracking System can basically be described as big data databases. Although in terms of volume they are behind scientific databases, i.e. data are not updated every second, but it is not so necessary anyway. Databases detecting weather conditions and storing their records contain more specific data since their characteristics can be measured even every second: temperature, humidity, wind intensity etc. In the meantime, such detailed assessment is unnecessary in career tracking – it is enough to collect for instance labour market data on a monthly basis, or student status changes by semester and not every second. The amount of stored data and their process is in accordance with the 3V big data characteristics: in the database generated in 2018 within the frames of a data integration procedure we store approx. 604.000 persons’ data, and collect more than 150 types of information from six administrative databases. Thus, the data amount we generate is over 2.5 billion single records. With such a population and in case of so many variables examined, it is necessary for one to be equipped with both IT skills and social knowledge to manage the whole research process. It is at a skills level similar to managing a big data database. Data integration and processing require special IT-like competence as well, interpretation of data requires a deep understanding of all the administrative databases and the collected variables, moreover, each research phase requires the understanding and acceptance of the big data way of thinking – with both its pros and cons.

After the formatting and cleaning of databases we can communicate data or do traditional sociological analysis in them.
Research modules of Graduate Career Tracking System

Data integration research modules
The data integration modules of the graduate career tracking system are based on the anonymised collection and analysis of personal information stored in administrative databases, and not on querying persons. The graduate career tracking system has two data integration modules: GCTS IAD (Integration of Administrative Databases) and the GCTS HECT (Higher Education Career Tracking).

Questionnaire research modules
Research modules based on online questionnaire queries are suitable for revealing relations between studies and labour market presence, charting students’ labour market expectations, fresh graduates’ status and position in the labour market, and also for exploring attitudes, opinions and motives related to employment, further national and international studies, and national and international work. The online questionnaire query consists of two sub-modules, namely, the Fresh graduate research carried out with graduates, and the Student research with students having active status.

The research started ten years ago in Hungary, and approx. 35-38 higher education institutions participate in data collection per annum. Educational Authority compiles a standard questionnaire, which is amended by questions that are of interest of the higher education institutions. Educational Authority provides the query software for editing the online questionnaire, through which the research can be carried out. Higher education institutions directly send their questionnaires to graduates and students, they also manage the whole research process, and then, on closure of the research, they submit answers given for the standard questionnaire supplemented by institutional information to the Educational Authority. After the participating institution send the responses of the questionnaires, Educational Authority arrange them into a research database, thus a national career tracking database is compiled.

Integration of Administrative Databases (IAD)
GCTS IAD data integration research module does tracking on the basis of graduates’ and drop-outs’ labour market data in a given year.

The IAD module of the graduate career tracking system does graduates’ career tracking analysis on the basis of factual data. In the last data integration, six data providing partners’ data were integrated. Different authorities (here data providing partners) provide different data: e.g. the National Tax and Customs Administration (NTCA) and the Hungarian State Treasury (HST) income data, data about
employers, information needed for calculating labour market status, the National Health Insurance Fund (NHIF) provides information about working abroad, and the Ministry of Finance (MF) provides data about job seeking and occupational trainings.

Conceptual model

The conceptual model of the GCTS IAD has renewed completely within the frames of the Human Capacities Development Programme (HCDOP) 3.4.5 Higher Education Sectoral Programme Aiming at System Level Development and Widening Accessibility project, and as a result, in 2019 data integration is done on the basis of the renewed model.

In the original model, the population consisted of those ‘graduates’ who obtained a provisional certificate of graduation within 1 and 3 years after completing studies. During the review it became obvious that the two series of the separate years are not suitable for real tracking. Thus we expanded the length of investigation period for seven years and we also enlarged the population: besides students with a provisional certificate of graduation, we included students registered as drop-outs of the given year, too. The population is defined on the basis of data stored in the Higher Education Information System.

Before 2018, the requested variable system was completely unstructured; duplicated set of data, illogical data were quite typical in the tables. Within the Human Capacities Development Programme (HCDOP) 3.4.5 Higher Education Sectoral Programme Aiming at System Level Development and Widening Accessibility project, we reviewed the requested variables and already structured them during the requesting process (the scale and nature of alteration is the biggest in the set of data from the HEIS).

Several relevant sets of data were missing from the data integration, e.g. besides the employed status there were no available data for analysing entrepreneur statuses. During the reviewing process we expanded the requested set of data, and also we aimed at the highest data quality within the categories, which was only possible if we requested further data (e.g. data of working abroad are far more improved and expanded in the new methodology).

Before the review and renewal, sets of data were requested from different sources (e.g. income), but the – in some cases quite significant – differences were not revealed or analysed, and the usage of source data was rather incidental. During the reviewing methodological workshops we identified and analysed the reasons for the differences, and standardised order of recalling source data to an analytical level.
The review resulted in a far more streamlined, focused and easy-to-handle dataset – bigger population, greater and more precise information besides lower costs (price of data integration decreased by 30%).

**A GCTS IAD core variables as a result of renewal by information provider partners**

**NHIF:**
- since 2018, in order to avoid duplication, we have not requested marital status and nationality as the latter is available in HEIS
- we do not request the table of employment data – they are provided by NTCA
- however, we do request data of living or transfer abroad in a separate table

**Ministry of National Economy (MNE):**
- we do not request data of gender and date of birth as they are available in HEIS
- data are accessible at a district level countrywide

**HST**
- we added data of earnings preceding 2012 to earlier data requests. Only data of total annual earnings are accessible before 2012 – these can be divided by the number of those days when the person in question was employed.
- from 2018, we differentiate pension contribution and income object to pension health insurance contribution, because while pension contribution has an upper limit above which there is no obligation for contribution, in case of the second one it is not the case. A significant difference between them is that pension contribution is paid by the employee, while pension health insurance contribution is paid by the employer.

**Student Loan Centre (SLC):**
- we request Student Loan 1 and Student Loan 2 baseline data from SLC, and disregard the other tables, e.g. we do not need data on repayment of instalments

**HEIS**
One of the most important innovations is that only three tables are requested instead of the earlier 14 – this enhances simple managing and requesting. On the contrary, it does not mean a decrease in terms
of datasets, moreover several datasets are included in requesting which were not before (e.g. language of study programmes). The three tables are as follow:

- Personal data – containing the personal data of the reference persons
- Data of courses – containing the reference population in the reference period and after
- Period table – containing semesters belonging to data of study programmes, thus we can track e.g. active/passive status, financing on a semester basis

**Block 1: personal data table**

Main changes in comparison to earlier data requests:

- Renaming columns; column names do not need to start with GCTS combination, student data start with HLG initials.
- If a student has valid record in more institutions, then the basis of the personal data block the particular record where the last update (HLG_LASTUPD) is the latest one.
- We define whether to include a student in the target group by OSAP repayment calculation, which is accessible in HEIS anyway, so it requires no further calculations.
- We expanded personal data by including district, county and region of students’ place of birth. These data are also accessible in HEIS, thus these do not require further resources and are suitable for geographical analysis.
- A new data is whether a student holds a card of Hungarian Nationality (Magyar Igazolvány). It aims to identify foreign students of Hungarian ethnic from outside of Hungary; they are not Hungarian citizens or dual citizens, but considered as Hungarian in the analysis.
- Other new dataset is the favouring of particular students, which may have added value in social dimensions later on. Five new columns are included in data, with (1/0) figures in each case:
  - any benefit a student is entitled for (orphan, half-orphan, breadwinner, large family)
  - disabled student
  - student takes care of one (or more) child(ren)
  - disadvantaged student
  - student with multiple disadvantages

**Block 2: courses table**

In earlier data integration there were more tables for obtained documents and language exams. We removed these two tables, reducing significantly the number of columns, we do not request for unused columns; consequently, we end up with fewer tables among the data which are easier to manage.
Fundamental innovation is that we request all relevant study data instead of requesting by academic years, so we have all data arranged in one table in the end.

In data request, the baselines are the course item marking the graph and the available courses, and not the course items of the given academic years.

We include all study information with the starting and the final dates of the period in one table instead of displaying all study data by semesters in separate tables.

Data refer to the registered courses and the programme, and their features: schedules, level, and language of the programme.

Place of the study programme is an important factor in the analysis, so we included in the requested data (with reference to the GPS coordinates, district, sub-region, county, region).

We included the type maintainer of the educational institution, as we find important (it was not included earlier though).

We aimed to handle the question of institutional succession as well, and accordingly, for succession and the data owner institute of the training we display the last known institution and faculty.

Besides the study field, we have data of trade groups and scientific fields, and a column with reference to pedagogical modules – in order to do a deep analysis.

In terms of dates, we have the date of beginning and completing the course, moreover, the dates of the provisional certificate of graduation, the outcome exam, the language exam and of obtaining the degree.

Later the following information can be of great importance: credits obtained, transferred or recognised.

Just like with credits, we request the number of semesters from HEIS: semesters completed during the course, active/passive semesters, number of (not) funded semesters.

We access state funded position of a course through the following logic: if the number of funded semesters is greater than or equals to the number of not funded ones then it is regarded as ‘funded’; otherwise ‘not funded’ (1= funded, 0= not funded).

We follow OSAP logic for analysing the course leading to the first higher education qualification.

Data include information on dual trainings, which can be an important reference factor for educational policy analyses.

Another fundamental innovation is aggregated information among training data on student allowances, scholarships, and state scholarships (Have they received any ..... for at least one semester during the training?). The impact of different funds, state scholarships, or even labour market significance can be followed up through these pieces of information.

**Block 3: Period table**
Period table contains data of training periods, semesters stored in the courses table. Significant data of the trainings are available here, as well as different features (status, method of financing) of a given semester broken down by semesters. The number of funded semesters, active/passive semesters, and methods of financing can all be tracked by these data both at individual or training levels.

**Method for data checking**

On the contrary to previous method of data cleaning by hand, we do detailed data profiling to check data quality instead, an also send feedbacks to data provider partners who have not received such information before.

**Data process methodology**

Over the traditional sociological data processing programmes (e.g. SPSS), the amount of data stored in data integration required an innovative methodology for data process. Data integration from 2018 stores far more than 2 billion of individual data, which amount cannot be processed by traditional sociological data analysis methods – there was a need for setting a standard methodology. We have voted for SQL, R and PowerBI solutions for processing data in an innovative way.

**Data visualisation renewal**

We have laid new foundations for data visualisation: it is done by PowerBI programme.
The Fresh graduate research has been carried out for nearly ten years in Hungary, thus it is considered to be one of the oldest systematic higher education data collections. When developing the research questionnaire and methodology, there was a completely different data collection environment for developing methodology and implementing. The next generations’ innovative use of media, dynamically improving technological changes made possible – and necessary – the overall renewal of the contents and methodology of research module. The reviewing of the graduate module of Fresh graduate research has become inevitable due to the growth and regularity of the data integration research module. We concluded as follow: within the project, an overall modification and renewal of the research module was necessary. The changes in data collection environment and the graduates’ questionnaire response habits, the expansion of the career tracking system lead to the fact that by 2018, the Fresh graduate research was not the only information source containing career tracking data; in other words, there were optional sources for answers. Moreover, there are questions to which data integration can provide better solutions (e.g. remuneration, work responsibilities). Data integration as a method will never be suitable for analysing attitudes, motives, and presumptions: only questionnaires are. Emphasis must be shifted towards these in case of the Fresh graduate questionnaire, and those questions/responses based on facts and data should be left for the data integration modules.

The most important step in the Fresh graduate questionnaire renewal was to modify the structure of the questionnaire for a modular one. This change made it possible that certain topics can be investigated every 3-5 years, but in more details and aspects. Before the renewal, we requested information annually about the graduate’s studies abroad – but never asked about its type, their experiences, if they found it useful for their further studies or work, etc. The direction of the change is to do a deeper analysis of key areas frequently, instead of requesting one or two responses on several different topics annually.

The reviewing workshops during the implementation of *Human Capacities Development Programme* (HCDOP) 3.4.5 *Higher Education Sectoral Programme Aiming at System Level Development and Widening Accessibility* project proved that the student module also needs modifications since there is no added value of working with exactly the same question sequences during the active study period. As a result, first the methodology for the Fresh graduate research (survey with students completed their studies) is modified (in 2019), and then comes the redesigning of the research with students having active status (in 2020).
Introduction to the research

In Hungary, Graduate Career Tracking System (GCTS) questionnaire module has applied basically the same research methodology for collecting data about Fresh graduates’ status and labour market position since 2010. An online questionnaire is used for data collection. At the moment, higher education institutions participating in the national programme of GCTS approach students/graduates through their own lists of graduates’ addresses and request data on the basis of the central research methodology provided by the Educational Authority Higher Education Analysis Department. The research questionnaire is accessible through a link sent to the participants via e-mail and to be filled out online. Questionnaires may contain institution specific questions besides the standardised blocks of questions, however, the national database only stores the standardised data property of the standard questionnaire items. Due to the institutional changes and other higher education institutions joining the programme, the number of higher education institutions participating in the research programme constantly changes between 30–38, whose students cover the total number of Hungarian student population with a 90% rate. Data collection is scheduled for the first half of the year (between March – June, during a few-week timeframe picked by the institutions themselves). Anonymised responses to the online questionnaire are collected and stored by the higher education institutions, then institutional databases – standardised by an interface specification – merged, cleaned and weighted by Educational Authority Higher Education Analysis Department on the basis of the scientific field of the qualification, course schedule, respondent’s gender and the year of obtaining provisional certificate of graduation. The statistical data of the population necessary for weighting is from the Higher Education Information System - HEIS. As a way of data validation, some responses to the questionnaire (year of obtaining the provisional certificate of graduation, study field, level) are provided by the institutions; they use data linked to specialisations of different course lists which are sorted according to HEIS and published on Felvi.hu (website of applying for higher education courses in Hungary). Having mentioned manual data entry, we need to take inaccuracies into consideration. The present procedure takes place at institutional level, and the national database is built upon their merger and integration. Until 2018, the target group of the research are, on the one hand, the active students every year, on the other hand, in case of the graduate research those having obtained provisional certificate of graduation 1/3/5 years earlier at the following levels: traditional university, college, Bachelor’s, Master’s and one-tier, long-term (5-6 years leading to Master’s) programmes.

System elements reflecting upon the needs for a reforms:

- participation (joining in and co-operation) in the research is optional for institutions
institutional reorganisations cannot be followed up precisely
- due to manual entry, institutional data providing within the research can lead to incorrectness
- responses in the standard database stored by Educational Authority are provided by the institutions, and not directly by the target group

Contents and structural renewal of the core questionnaire
- contents renewal (revision and updating of questions, set of questions)
- structural renewal of the research (e.g. central managing and application of register based survey solution instead of FOI request).

Contents renewal
A contents renewal of the research, the restructuring of the questions in the questionnaire has been an ongoing process for years, which meets the needs of higher education institutions as well as the methodological renewal concepts of the Educational Authority; and expectedly – considering the decreasing number of responses – with the young people’s needs just as well.

As a result of reviewing the Fresh graduate research module, the following general principles were concluded:
- renewal of the research concept
- reducing the number of questions
- simplified wording of questions
- logical overviewing the order of questionnaire blocks
- reducing the number of branches

1. Renewal of the research concept
In the present system, every graduate receives three times a questionnaire related to graduate career tracking, which is rather demotivating due to the frequency and the same sequence of questions. Furthermore, those doing further studies at a Master’s programme after their Bachelor’s and complete them in time will receive two questionnaires at the same time 3 and 5 years after graduating at the Bachelor’s programme (and 1 and 3 years after the Master’s) from one or even more institutions.
According to the renewed methodology, every single graduate receives a questionnaire from the institution 1 and 5 years after graduation (dismissing the query of the third year). Why we vote in favour of the first year is:
- e-mail account database is more reliable
- more respondents
- graduation is closer in time than when we request data e.g. 2 years after graduation
- ensures continuity with earlier queries

2. Compiling various questionnaires

The main aspect in the renewed methodology of the Fresh graduate research is still differentiating the questionnaires on the basis of the time passed since graduation, and not the level of qualification. However, we would approach graduates with questionnaires of different focuses 1 and 5 years after graduation.

From a professional point of view, compiling varied questionnaires according to the time period since graduation is more desirable because respondents are at different stages in their lives at the time of questioning, and the questionnaires need to reflect upon this fact. By the fifth year after graduation, respondents will have gained such experiences – besides higher education studies – that are not so strongly related to their higher education or particular institution(s), but are of much significance to us when analysing labour market life paths. Five years after graduation, young adults usually have quite a stable work position, so analysing and identifying the most important factors (especially further studies leading to that stage) can serve as a feedback about higher education study programmes to both educational policy and the institutions (e.g. to what extent could the respondent benefit from what they learnt during their studies, did they need to participate in any further training).

Fresh graduate research is implemented with modules of different scopes for graduates 1 or 5 years after graduation.

Short term feedback (1 year after graduation) research

Research solutions based on short-term feedback have more existing models today. In Australia between 1972 and 2015, the Australian Graduate Survey (AGS) queried former students 4 months after graduation. In Ireland, since 1982 the First Destination Survey career tracking system has been carried out annually through approaching student shortly after graduation (e.g. 9 months after graduation for those graduated in 2017), and involving every single state university.

Considering experiences of countries having short-term feedback questionnaire (too), the main focuses of the questionnaire are as follow:
- current work status (nature of status, income, one or more jobs, nature of the job and the organisation, how long they have been working there)
- the process of finding current job and the reason for choosing it
- employment during studying
- unemployment
- satisfaction with the current job, any intention to change
- connections between higher education and the labour market position (fitting, gained and applied knowledge and skills, knowledge elements)
- general feedback on the quality of the courses, study programmes for the institutions (if we can find mutual elements in the earlier institution specific questions)
- current studies (e.g. language course, vocational training – including higher education studies from HEIS) and future study plans
- studying or working abroad, or any related future plans
- personal data (socio-demography, address, family)

Mid-term feedback (5 years after graduation) research

International researches have proved that 4-5 years after graduation such information can be accessible about labour market positions that may forecast not only the early stage of a career path but a mid-term success, too. In the Labour Market Career Survey of Young Graduates (Diplomások Életpálya Vizsgálata – FIDÉV) research implemented between 1999-2004, worked with similar time periods, and the responses proved that graduates’ positive early labour market position further improved during their life paths, and their labour market integration better reached its full potential. Research findings highlighted that further studies (regardless if it was formal or non-formal) played a significant role in young graduates’ successful labour market integration.

The query scheduled for 5 years after graduation can contain more items on labour market success and factors having an impact on it. Main topics are:

- current work status (nature of status, income, one or more jobs, nature of the job and the organisation, how long they have been working there)
- reasons for current position/job and multiple employment
- motives, career plans, satisfaction with the current job
- connections between higher education and the labour market position (fitting, gained and applied knowledge and skills, knowledge elements)
- history of employment (features of first job, length of employment/unemployment altogether), especially business experiences of entrepreneurship
- formal and non-formal studies in relation to professional development, voluntary job, etc.
- studying or working abroad, or any related future plans
- personal data (socio-demography, address, family)

3. Modular questionnaire

Filling out the current questionnaire is rather challenging. Too long, too many branches, which make managing the questionnaire and data analysis difficult, moreover, its use is limited due to the low number of responses to conditional questions. Too long, thus adding institutional questions also narrow possibilities; respondents often claim the questionnaire is boring and undifferentiated – each group is given the same questionnaire each year.

The renewal aims to shorten the questionnaire – standard questionnaire modules of the Educational Authority should require a maximum of 10 filling requests even in the longest branch. Containing just a few branches, being simpler, more obvious and trendy, and involving institutions added-on questions are also among the biggest goals.

*Forming a modular structure*

Modular structure means that the questionnaire consists of different thematic blocks (modules).

Modules:
- core questionnaire module: the same in case of each query
- differentiated module: diverse modules in questionnaires for graduates 1 and 5 years after graduation. One year after graduation the focus is on the transition period, while five years after graduation the focus is the labour market life path.
- thematic block: a module aims to enhance a deeper analysis of particular topics (in comparison to today's practice of one question involving more topics, thus a thorough analysis is rather problematic e.g. working abroad). Thematic modules change annually, focusing on particular topics each year – and we make requests on the same topics every 3-5 years.
- institutional block: questions compiled by the institutions themselves.

*Structural renewal*

Within the frame of the project, in 2018, a pilot research was conducted. As a result, we have developed an innovative research methodology unique in Europe: the register based survey research. Within the
pilot programme, we connect in an anonymised way administrative data stored in HEIS to questionnaire responses of those graduated in 2018. For implementing the research, a unique IT development is necessary; the modification of the data collection has a great impact on methodology, too.

Register based survey unites data integration and methodology for questionnaire data collection, thus its delivery requires the central management of Educational Authority. Within the pilot programme, as for data collection, we changed for the centralised Educational Authority research, so data collection and analysis was directly delivered by Educational Authority Higher Education Analysis Department (structural renewal).

System elements proving the needs for the register-based survey

- participation (joining in and co-operation) in the research is optional for institutions: Article 25 (1)-(4) of 87/2015. (IV. 9.) Government Decree
- institutional reorganisations cannot be followed up precisely
- due to manual entry, institutional data providing within the research can lead to incorrectness
- in the standard database stored by Educational Authority are provided by the institutions, and not directly by the target group

Register based survey data collection

The source of the respondents’ email accounts is HEIS, since the sectoral minister is in the position to request any research based on email accounts stored in HEIS from the Educational Authority. As a result of the structural renewal, the change for administrative bases reduces the importance of institutions in implementing the centralised research, in the meantime, it provides the possibility of requesting information about not only the graduates/students of institutions participating in the research but about the total target group, too. The primary aim of the centralised survey is to set a national sample.

Higher education institutions still have the possibility to compile a unique block of questions typical to their given institution besides those in the standard questionnaire. Responses given to theses institutional blocks are sent back to the institutions by the Educational Authority. Thus the direction of data flow changes: up till now, in accordance with applicable legislation FOIs carried out the research and sent the responses supplemented with institutional data back to the Educational Authority – the direction of dataflow is the contrary in case of the register based survey research. For FOIs the advantage of the modification is that institutions analyse the returned data base instead of having the duty to collect data. This basically means that human capacity of the research (groups) within the institutions does not need to be reduced since all tasks except for data production remain with them.
Data collection was carried out in 2018 by Evasys programme with an updated function, anonymity is ensured during the requesting process. Educational Authority receives an anonymised data base after the research closure. Anonymisation would be ensured in the questionnaire software system, thus inserting and linking administrative data as data table with the questionnaire response data happens by anonymising the connection code at the moment of submitting/sending. After connecting inserted administrative data with responses, the software deletes connection code straight away thus personal identification on the basis of result database becomes impossible.

Statistically speaking, register based survey method produces more valuable, user friendly and modifiable databases. Further advantage of administrative data usage is that it reduces respondents’ “workload” since information is only required about those data missing from the administrative databases. This leads to producing shorter and user friendly questionnaires meanwhile it opens ways to do analysis through involving the most background variables possible. Questionnaires must not be overloaded with such administrative data as the number of background variables, because there will be no time left for questions of real interest. Apparently, this method does not require requesting background variables (these are included among information stored in administrative databases) since higher education students’ or graduates’ educational/socio-demographic background data can be directly derived from HEIS operated by the Educational Authority. Thus more possibilities are left for revealing attitudes and studying opinions.

See flow chart of the process below:

Detailed data of the target group from HEIS require data cleaning and producing of aggregated variables before applying them in the questionnaire research because we can only derive raw information from the database – more complex datasets, defined in compliance with research interests are not available in HEIS directly (calculated indicators); naturally, since its primary aim is to store data of education and training, and not a research.
One-line approach
The course evaluation question sequence of Fresh graduate research has shed a light on a crucial conceptual and research designing issue. 5% of the respondents may receive career tracking questionnaires of more courses at the same time. IT principle is to find solution for even one single case – while 5% of multiplicity was rather high in the piloting year. We need to provide possibility for the respondent evaluating each of their courses (we met only one extreme case: in 2018, one particular person was registered in four courses in HEIS as a graduate).
We changed the previous research concept of educational level for a one-line approach instead – uploading all data of the courses data in advance under each respondent. During the responding process it became obvious which course the respondent is referring to, and we also made it possible for them to respond to the questions of any (or even all) courses.

Results of the methodological renewal
Opportunity for handling institutional changes
The frequent and large changes in higher education structure imposes burden on implementation at an institutional level as there is no general methodological solution or best practices for the career tracking to follow structural changes in a given institution anyhow. We cannot expect, as most often it is impossible anyway that an institution tries to track graduates who literally never were their students.
On the other hand, we cannot expect either that graduates of faculties transferred to other institutions are still approached by the institution after the transfer; in other words, institutions are obliged to contact only their own graduates. Institutions do not apply a standard methodological solutions but a provisional one for tracking student groups involved in such transfers, while they are typically not included in the national database. Institutional changes generated an increasing blind spot in earlier data collections in terms of the survey.
By 2018, within the HCDOP 3.4.5 project, the matter of tracking institutional successors in HEIS as an administrative database was solved, thus we can follow institutional reorganisations – however, the transition for one-line methodology solves the problem of succession anyway.

Shortening the questionnaire
It is a recurring issue about the questionnaire that it is long and diverse, which tends to discourage respondents. Duplications are a problem, too (target groups are formed by courses): a student receives the questionnaire as many times as they are selected for the population on the basis of different criteria (active student, or graduated 1/3/5 years before). Besides these, in case a student attends/attended
courses at different institutions, again more institutions may approach them with their GCTS questionnaires. These factors all lead to the overload of potential respondents, which altogether decreases enthusiasm and systematically biases results because respondents are quite unlikely to respond to latter questionnaires.

As a result of the innovation, overload is ended since the student is given only one questionnaire, which refers to each of their courses, and it is up to the student which course they want to respond to.

Pilot research technical data
The pilot research was implemented in the autumn of 2018. Respondent rate was higher than 20%, which exceeds respondent tendency of the earlier Fresh graduate research. It overgrows our expectations as well, since at the beginning of the research we did not have information about how efficient the Educational Authority might be in mobilisation. Unlike the institutions, Educational Authority does not have personal, long term and close relationship with graduates.

Respondents’ rate
- The pilot questionnaire was sent to 35711 persons
- Bounced messages after the first sending (failed delivery): 1332 (3.7 %)
- Number and rate of respondents: 8350 persons (23.38 %)

Results
The pilot research proves that responds can be technically connected to data stored in HEIS. Connection tests and validations in EvaSys show that the register based survey method is successful; however, data integration process required some months and continuous IT support to be implemented and managed.

Developing the methodology for the register based survey within the project enhanced doing thematic researches, set methodological basis and outlined the necessary IT background for the implementation.

Higher Education career tracking
The project aimed to enhance research environment for analysing input factors besides the outcomes in career tracking. Methodology designed especially for the Higher Education Career Tracking project enables parallel analysis of higher education and labour market life paths. This methodology can be applied generally in every field of education. We implemented data integration, set up a pilot research database, tested methodology and analysed data in one particular educational field, namely IT in the Human Capacities Development Programme (HCDOP) 3.4.5 Higher Education Sectoral Programme Aiming at System Level Development and Widening Accessibility project.
New research methodology

Higher Education Career Tracking (HECT) is a data integration research module similar to Integration of Administrative Databases (IAD), however, its conceptual model is different. IAD focuses on graduates of a given year while HECT follows what happens to an entering class during the course period and then in the labour market. Due to different methodology of the research, different research query can be carried out with different assessment and analysis focuses. IAD focuses on completed courses of a given year and HECT follows the life paths of courses beginning in a given year. See the following example for representing the difference:

- IAD examined those labour market positions who dropped out or graduated e.g. in the 2015/2016 academic year. It can differentiate the market labour position of graduates and those without a degree at the time of the query (employed or not, work in a graduate position or not, have higher or lower income, etc.).

- HECT focuses on what happens to different courses during a given period, e.g. if a course programme starts in the 2010/2011 academic year and lasts till 2015: how many course programmes are completed successfully (i.e. with obtaining a degree), or terminated due to course change or suspending studies – and what happens to dropouts’ life paths or that of those changing courses during their studies. Thus it provides information about a complete entering class on the contrary to IAD focusing on graduates – it is irrelevant if graduation happens after 3, 5, 8 (or in an extreme case even 24) semesters. All in all: HECT examines a complete entering class and able to analyse parallel study paths in case of those who graduate as well as of those who do not; IAD focuses on graduates of a given year – and regardless of the date of beginning their courses, it ‘mingles’ students of several different years.

Professional and technical dimensions of methodology:

Professional dimension

We considered the following principles during the developing of the methodology:

- We require every single data from the same source, avoiding duplications this way
- We do not require irrelevant data
- We define the necessary datasets in accordance with the researched questions raised
Time horizon of the analysis: from July 2009 (2009/2010 academic year) as long as we have available data, namely the 2015/2016 academic year we examine the entering classes. Consequently, the population consists of students entering higher education IT studies in the 2009/2010 academic year, or later – regardless of the outcome or completion of their studies.

Through this wider population and more years of data selection, we can analyse the labour market position of those entering their courses (even in yearly breakdown), moreover, after this, by querying similar population, we can track every year the entering classes even in a seven-year timeframe – this enhances a dynamic follow up of graduates in each class, which can be important and informative for educational policy as well.

**Technical dimension**

**Methodology of data connection**

Data connection method is in compliance with Integration of Administrative Databases module method, National Infocommunications Service Company Limited by Shares (NISC) by law anonymise personal details in data tables by law, which enables connecting tables. The linking filed between the tables is the Social Security Number (SSN). Other data providers also use Tax Identification Number (TIN) or SSN (depends on the data provider), so data received from other institutions can be linked by these fields (linking fields).

Hash algorithm applied during data connection is run by every data provider on their own data, so NISC only sees encrypted data. After NISC received all required data, NISC produces an increasing sequential number code form the hash codes. This way it is impossible to decrypt the code and it is in compliance with data protection regulations. See the following flow chart for methodology of data connection:
The most significant element of technical renewal is that we transfer data into an analysis database, where we check and clean data more times. During data process, we connect data, the necessary dictionary and the core tables, e.g. meaning of SCO (standard classification of occupations) codes, etc. This data process results tables which can be efficiently used for analysis and providing any kind of data.

**Variables involved in the research**

Higher Education Information System data

- personal data in HEIS: data connection is based on this. Data are from HEIS and contain the most relevant personal information.
- educational data in HEIS: detailed data about the study programme (institution, major, schedule, date of beginning/completing the course, performance, dropouts, etc.)
- data about academic terms in HEIS

Hungarian State Treasury (HST) data

- data of workplace and employment

National Health Insurance Fund (NHIF) data

- data of social security and staying abroad
Ministry of Finance data

- data of job seeking register and active labour market tools

Research findings – career tracking of IT educational sector students

Higher education career tracking research module has a dual aim within the project: besides developing methodology, it also aims to analyse career tracking data of IT educational sector students.

Analysis aims to reveal what a difference the IT education means in the labour market in case of dropping out, completing the final exam (with leaving certificate) or graduating. Thus the most significant criteria for selecting the items of this population is students entering an IT study programme, and then we collect information about them from other administrative systems.

The analysis mainly focuses on the difference between dropouts and graduates on the basis of the following criteria:

- labour market position (SCO, income, active/inactive labour market status, child care allowance (GYED), child care benefit (GYES), job seeker)
- length of time spent with job-seeking, working during studying
- dropout-BSc-MSc – studying, further studies, refresher courses

Most significant figures of data connection are:

- 34501 students
- 66644 programmes
- 364572 period of training

In the research within the HCDOP 3.4.5 Higher Education Sectoral Programme Aiming at System Level Development and Widening Accessibility project each examined study year can be analysed separately or in a complex structure. Now we introduce in a nutshell the most important data of Bachelor’s and Master’s entering classes of the 2009/2010 academic year because this is the longest period (7 years) within the research – it is already a mid-term tracking and as such suitable for drawing well-based educational policy conclusions. Below we introduce the most typical prevalence for analysed population, and then the research findings. The number of persons in the population is 4822, which means 4842 courses.
All in all we can conclude as follow for those entering IT educational sector in the 2009/2010 academic year:

- Dropouts rate is rather high, especially in first cycle programmes of higher education.
- In terms of income, there is a significant difference between graduates, those obtaining a provisional certificate of graduation and dropouts, in other words labour market prices dropping out and the degree as well.
- Dropping out correlates with the amount of work one does during their studies.
- There are regional differences countrywide regarding income.

**Dissemination**

The results and findings of graduate career tracking are regularly published by the Educational Authority in *research reports, annual and sectoral analysis*. Since they are informative, career tracking data can support developing sectoral strategies and educational policy decision making.

Research data support *career plans* and *career guidance* also: info graphics of educational fields and majors can help orientation for the secondary school student target group, which we display at the beginning of the enrolment procedure each year as part of an introduction to the given major.