Perspectives and problems using administrative data for labour market monitoring

The digitalisation of information in society is leading to an entirely new phase in labour market monitoring and research – and in social sciences in general. More and more events in life are registered electronically. This digitalisation is just two or three decades old – and it is a process that is still spreading into new areas of our life: changing employer, passing an exam, buying a chocolate bar late night in a 7-Eleven, making a mobile phone call, paying the membership fee to your local sports club. The electronic registration of these events leaves a digital trace and all these digital traces make it possible for social sciences to get very detailed micro data – if they could be integrated. This of course creates fears of a "Big Brother Watching You"-society and that is an important discussion. This article aims to describe some of the possibilities, problems and perspectives of using all the digital traces.

The key to grasping the enormous potential is to understand what is needed to get an integrated register data system (IRDS). In a world of digital traces we are not longer talking of "data sources" as a fixed set, in the sense that the Labour Force Survey is one data source, the census another. Often data was lacking for social sciences – so one had to make a survey. We are not talking about a limited number of known registers, but about the totality of digital traces that can be integrated in order to answer our research, policy and administrative questions.

The two key identifiers: persons and organisations

Society consists of basically two types of actors: individual persons and organisations. If we have a unique person identifier and if we have unique identifiers for organisations as well as contexts where we can make a connection between these two, then the digital traces can be integrated. This integration means that social scientists will have hitherto unknown amounts of data and will come closer to natural sciences in some respects. It will no longer involve the handling of data deficiencies. Non-response for example is not an issue; the registers always "answer". Given an ocean of micro data, it is difficult to establish the best trade-off between detail and overview. Finding the most appropriate level of aggregation becomes the major issue, not missing data.

In the Nordic countries a unique person identifier was introduced in the sixties, in 1964 in Norway. This was in the very early stages of the computer age and was to a large extent a reflection of "paper-based" administrative needs. Another driving force was the introduction of big mainframes in some official agencies like Statistics Norway, the tax authorities and the social security agency. In a modern state, where people move around the

country and where banks operate on a national level, name and address is no longer sufficient as identification. A unique national identifier was one way to meet this need. Since this was before the general public's awareness of the possibilities that computers give to match data, there were few – if any – worries about privacy. There was no Internet, so computers were physically isolated. The obvious efficiency/consistency gains led to a fairly general use of the person identifier as time passed by. First of all in the public sector, but also in banks and insurance companies, in the personnel files of big firms – and it spread to non-commercial, voluntary organisations.

The myth of the Nordic centralised state

A common misunderstanding is that in order to have a national system of register data this has to be planned and maintained by a central authority. The OECD manual on Human Resources in Science and Technology (the Canberra Manual) describes it as follows: "Some countries, especially the Nordic ones, have a tradition of centrally co-ordinated registration of characteristics of individuals". But this is not correct. There was never a centrally co-ordinated registration of individual characteristics. It was the problem of uniqueness, i.e. that too many people have the same name, surprisingly many live in the same locality, even apartment (father and son, mother and daughter have the same name) that led public and private institutions to use the person identifier. The public sector disposed of databases built up over time which were not intended to be matched by the unique identifier or matched at all. But as soon as you store information using the person identifier, you have data that technically can be matched. You have what is called a "distributed virtual database". The metaphor should not be "Gosplan"[1] or the hyper-centralised state. The metaphor should be the "invisible hand": if everyone by selfish need of efficiency (uniqueness) uses the standardized, national identifier, then "order out of disorder" is created. Separate, fragmented data start "talking to each other" and can give answers to important administrative and scientific questions, either labour market related or related to crime or to the "war against terrorism".

On the contrary, the lack of centralisation – or rather the lack of coordination – is a problem. Public authorities are not aware of the "force" of joining data. The central population register is frequently used: surveys do not ask people about age, sex or where they live. The respondents just fill in the person identifier and a computer finds their age, sex and address in the central population register.[2] The business register is clearly under-utilized in this respect. Both public authorities and researchers ask about address, total sales, export, number of employees, number of women among the employed, educational level of employees etc. – making the firms irritated, complaining (correctly!) about bureaucracy. All of this information they could get from the business register. It is as simple as that.

The use of a national, standardised, unique organisation identifier is much more recent, from 1995. For decades we had several identifiers:

a) The "statistical" firm identifier used by Statistics Norway;
b) The "social" security firm identifier used by the public social security agency;
c) The "VAT" identifier used by the tax-authorities.

It is beyond the scope of this small article to discuss why this chaotic situation was not corrected before 1995. But the basic explanation is the interaction of three processes:

a) The rising need for integrating data, for re-use of already collected data;
b) The increasing practical possibility of integrating data and the growth of networks;
c) The collective "system" awareness of the need and the technological possibilities.

There was a government green paper in 1988 that outlined a unified business register, but it was not until 1995 that the system was up and running. The system consisted of two firm identifiers: one for establishments (local activity unit) and one for enterprises (legal, owning unit). This was clearly a "social construction of identifiers". Why not three layers: establishment, enterprise and multi-enterprise corporations and holding companies? This is now being constructed – since there is a need for it. One could have had four identifiers, the fourth being a number for the "profit-centres" inside the establishment.

It is important to keep in mind that the "business register" easily gives the impression that it only includes commercial firms. But this is certainly not the case in Norway. All public institutions (state, regional, municipal administrative authorities, all schools, hospitals, prisons – i.e. all public employers) are part of the register, as well as a wide spectrum of voluntary organisations, with or without employment. In Norway you have to be in the business register if you want to have a ".no" domain on the Internet. This means that a lot of organisations that currently do not have employment – but potentially can have – are in the register. This is very useful for labour market research and other types of social science research on cultural or political issues. Another important organisation is the family and/or the household. With register data one can construct families and households, using various digital traces (birth data, adoption data, address information, martial register etc.). Parents can be related to children, married couples as well as legally accepted informal marriages (having children together and living at the same address).

The crucial nexus between persons and organisations

Using unique identifiers for persons and firms is not enough for most purposes, and especially not for labour market research. To get an integrated register data system, we must be able to match employer and employee. In Norway this was done using the social security register, which in principle should be updated every time a person is employed by an employer (and each person might have several employers) and each time the employment relation ends. The rationale of the social security register is payment of benefits (long term illness, disability etc.). This means there is actually more information on the "sick" since monetary transactions (digital traces) are involved, while for the stable employed, the more normal, the higher educated, less data are monitored by monetary transactions. The alternative to using the social security nexus was to use the tax authorities. In that case the data would have been better screened, both by employer and employee since "big money" is involved. But still employers do report taxes on an *individual* basis only once a year for the tax declaration.[3]

This illustrates the general rule that when there are several sources for the same data, one should choose the one that is best screened on quality by real-life events. Taxes concern everybody, benefits only those that receive them and both employer and employee check the data when sending/receiving the pay-check. Consequently, in connecting employer and employee one should use the tax-data. Statistical collection is in a certain sense an "event", but it is not an event that is crucial to either the employer or the employee, so they are not directly, immediately materially interested in the correctness of the data.[4]

Statistics Norway should more and more use the digital traces instead of self-made surveys. Both in the short and the long run, this results in more, cheaper and better data. The Danish and Finnish have since the early nineties stopped making the traditional 10-year paper/CATI-based census. Almost all important information is already in registers. A good example is that in Norway there is a clear tendency to use registers to fill out originally "paper-based" surveys like the Labour Force Survey. Once you have the person identifier you do not have to ask the respondents about education or last employer – you get that from the appropriate registers. You might ask this information in order to check people's memory or to check the quality of the data in the register, but basically that is not necessary. In the future one should not use sample surveys like the Labour Force Survey to get "hard" data (age, sex, marital status, income, address, highest education, previous jobs etc.). Sample surveys will be used purely to collect the *subjective opinions* of people. The "objective" part of the Labour Force Survey should be drastically reduced and the "subjective" part should be expanded.

The use of administrative registers in labour market research

Monitoring labour markets by use of an integrated register data system (IRDS) has been exploding the last ten years in Norway, since Statistics Norway made a basic matched employer-employee dataset. Time series started in 1986 – updated each year. Data for 2006 will be available in June 2007 resulting in a twenty year time series.

It is important to understand that we are *not* talking about a fixed set of variables delivered by Statistics Norway – the usual "take it as it is"-dataset. In an IRDS, data about people and firms can be constructed according to the analytical question at hand. For example, since education is often very useful it is a default variable, marital status is not. But if you need it – Statistics Norway will add it. Do you need to know which persons are married to analyse the interaction of spouses' activities with respect to the labour market (or housing market) – just ask for it and Statistics Norway will give you the information needed.

It is very important to have a creative, holistic approach to monitoring. The mechanisms of the labour market are conditioned by all kind of factors. If we think something is or might be important, we must creatively think about what kind of digital trace it might have left – and where it is digitally stored. Let me take one example from my own research. The Norwegian Ship-owners' Association wanted to know what happened to the sailors when they went "on shore" as a consequence of the replacement of high-wage Norwegian sailors with foreign nationals (low wage labour). Since some of them continued to work for the same employer, but not on an actual ship the default variables in the matched employer-employee data set from Statistics Norway, i.e. change of firm identifier from one year to another, would not tell who went on shore and when. But from the early sixties there was a special pension agency for sailors, that had two different payment regimes for "on ship" sailors and "on shore" sailors. The change of regime was registered on a monthly basis. Since the person identifier was used in both registers this crucial variable could be added to all the other relevant personal and business information. The "Sailor pension fund" register had been established without any intention of being used for monitoring the behaviour of sailors. The topic of high-wage (Norwegian) versus low-wage labour was a non-issue in the sixties and seventies.

This again illustrates that the key to an IRDS does not consist of planned, co-ordinated, intentional built-up data sources. The general rule is that every personal register of any kind uses the person identifier and that any register of organisations (firms included) uses the organisation identifier.

The "full count" aspect of registers is also very important. To take another example, if we want to study international mobility of Norwegian PhDs, we join the matched employer-employee data with the migration data and we can see how many of the PhDs that left Norway in a given year have returned/stayed in the US – to take one important destination. The number of female computer science PhDs going to the US is such a minuscule group that the Labour Force Survey would probably have no observations at all. The Eurostat and OECD are now implementing a huge, costly, survey of PhDs in order to try to answer such questions. With register data you could have answered the questions using a fraction of the costs in a fraction of the time. The examples of this type are endless and demonstrate that the future of social statistics belongs to register data. A lot of what is now labelled as "unobserved heterogeneity" will become observed heterogeneity when researchers creatively use all the digital traces. Another example is two persons educated as economists, one having taken courses in computer science. Very often social scientists only have a rough measure of the level of education. In Norway the default data is the level – and a fine grained (3-digit) code of field of study of the highest achieved education. But also available is every single exam passed after the age of 16 – including ICT-courses. Having ICT-skills certainly made a difference in the labour market in the last two decades, so detailed information about that could tell a lot about the extent to which skill mismatch really was or is an issue.

Let me again underscore that it is not only public and private business registers which are repositories of digital traces. The Curriculum Vitae could become a very important source for labour market research. Here we find data on competences like the knowledge of languages – both natural like English, German etc. but also programming languages (Java, C++). This is important data which is hard to find in any other single source, especially for those countries not having an IRDS. The big recruitment firms (Manpower, Stepstone, Monster) have large built up databases for matching employers' demands to job-seekers characteristics. There is an "HR-XML Consortium"[5] working on standardisation of such data and the related business processes. For labour market research electronic CVs already stored in databases can be very useful. For

example in getting data on what kind of skills are in increasing/decreasing demand, statistics on type of competences demanded by employers using these electronic CV databases would be very useful.[6] In order to move from the current "word-file" to a really structured, data machine readable document however, a large scale standarization is needed.

The problems of register data

Although there is a bright future for social sciences using digital traces as an extremely detailed data source, there are of course some problems that must be dealt with. The problems can be divided into four categories:

a) Problems of sampling versus event registration;
b) The social construction of events;
c) The zero-tolerance approach;
d) Technical pitfalls.

Sampling versus event histories

There is a very long, almost "natural" tradition to collect data on a yearly, quarterly or monthly basis and to use "snap-shot" sampling. For example, the number of employees is equal to the number on a precise (sampling) day. In Norway, the employer-employee data are sampled at the beginning of November each year. This means that all information about job changes or change of labour market status before and after this update is lost. For the stable part of the work force this is not a big issue, but when studying people with frequent job changes, i.e. the more "marginal" groups on the labour market, it is important. Not only does the sampling lead to underreporting job changes, but also the *causality* is often lost. Especially when looking at events like participating in labour market (training) programs, participating in "public work" schemes – the timing (causality) of events are of crucial importance. Fundamentally, all data should be registered with the most fine-grained timestamp possible. If this is done, the timing of events is not lost, resulting in the possibility to aggregate to the type of period suitable for each analytical purpose. An event based registration – as opposed to sampling – makes it possible to use event history analysis (duration analysis).[7]

The social construction of events

The definition of an event is basically a political question. In a democratic society, civil society can decide via the mechanisms of representative democracy that certain information should be accessible and, if necessary, created. Let me exemplify this with the question on economic (accounting) data on *establishment* level. A multi-plant (establishment) firm only reports accounting data on *enterprise* (legal unit) level. This means that some plants might be profitable, others run a deficit. People might be hired at one plant and sacked at another. If one tries to investigate the relation between profitability and hiring/sacking, this relation might be completely lost when you only have accounting data on enterprise level.[8] Another example is that public project support (R&D and innovation) is mostly given to the *establishment* (plant) – but again – data on economic performance is only on enterprise (legal unit) level.

Data on profitability in each plant are already produced to a very large extent seeing that multi-plant firms tend to regard each plant (and often departments in each plant) as profit centres since it is obviously important for the firm to know the profitability of its different plants. This means that in the accounting system there is data on each individual plant, but since the law does not demand data on establishment level, they are not available. But if we wanted to have access to these already existing data – or force the enterprises not producing these data today to do so – it is up to democratic society to decide. Or in other words, the data are physically there already, but they have to be *socially* constructed, i.e. made accessible by agreeing to grant access through the mechanisms of representative democracy.

The zero tolerance approach

When working with register data one often encounters quality problems and when complaining about them we get the answer that these data were not made for research but for administrative purposes. But this is to turn the problem completely on its head. For research purposes, some missing data, some "noise", is not a problem. There are well-known and tested procedures to deal with

that. For administrative purposes, where correct/equal treatment of each particular firm's taxes is important, the "noise", i.e. missing and inconsistent data is a real problem. There are establishments that do not have a NACE-code or that do not have the code of the municipality where it is located. But the definition of an establishment is an activity that is geographically localised[9] and has a NACE-code. There are not many of these missing values, but there should be none! In the case of the municipal code, there is often in the same database more than enough address information (postal code) to get the municipal code right. In the paper age one could not have the ambition to have completely consistent data, but this should clearly be the ambition in the computer age – especially for administrative purposes. But for both administration and research it would save a lot of time and frustration if one always knew that the elementary data where not missing. Norway has a "legacy" problem here, but anyone building up an IRDS today, should adopt a "zero tolerance" from the start. This is a central indicator of quality/consistency of statistical data based on an IRDS.

Technical pitfalls

As mentioned above – it was the social security register that became the starting point for matching employer and employees, but unfortunately they had not used different/individual numbers for enterprise and establishment. They used a MS-DOS filename like 8+3 number structure. That is, an eight digit number for the enterprise (legal unit) and three digits for the establishments. But what happened when establishments were sold among enterprises as a result of bankruptcy, mergers or aquisitions? All information on the establishment was lost[10]! Fortunately the new and unified business register uses two independent nine-digit number series.[11]

Another example showing the technical pitfalls is the change to the worse from ISCED-76 to ISCED-97. ISCED-76 was a compact 4-digit number; ISCED-97 is both letters and numbers. As could be foreseen, having to use several variables, i.e. the four digit level and field of study increased the workload and "destination", "orientation" and "duration" are now separate variables. Although ISCED-97 is more detailed, people actually use it on a less detailed level than ISCED-76. Most researchers aggregate everything to 5a. In labour market research this is regrettable, because even a small difference in educational attainment or in the field of study might change labour market behaviour considerably. "Compact" indicators are mostly better than "fragmented" indicators.[12]

The "Big Brother Watching You"-syndrome

The digital traces (registers) are produced in all modern, ICT-based societies, but access to this type of information is a very different problem. It is only the Nordic countries where we can talk about an IRDS. The major obstacle in most countries – besides organisational inertia – is the issue of privacy. This is of course a complex issue, but I think that the best starting point is to consider that for decades there has been use of register data in the Nordic countries and there have not been any scandals, any examples of misuse. Secondly – one cannot on the one hand want eGovernment – which demands a high degree of "back-office" integration of registers in order to serve the public in an efficient way – and at the same time being sceptic to the integration of registers. One cannot have one's cake and eat it.

One should also be aware that one is already watched by two different type of big brothers. Firstly, when people use "bonus cards" their shopping habits are registered in detail for purely commercial purposes. Secondly, the secret police already has, by formal/legal or informal means, access to a mass of electronically stored data like e-mail and mobile phone calls. When they need data they will get them. The computer based work of the German Verfassungschutz in tracking down Rote Armee Fraction is a well-known historic example. Since then, the possibilities have become even larger. The quarrel about what information European air companies should give to the US authorities on passengers going to the US only shows the tip of the iceberg. The Echelon system is another example. In light of this it becomes a bit ironic that it is only publicly accountable actors like state administration and researchers that do *not* have access to register data they need to serve the

public in a cost efficient way. The possibilities for saving administrative work are enormous and the potential for improved service to persons and firms are great if an IRDS is established and if access is given to researchers.

Anders Ekeland
NIFUSTEP, Norway

Notes

1. The State planning commission in the Soviet Union (udarstvennyi ovyi komitet).
2. The gender is given by one of the digits in the number, so no check is actually necessary.
3. Employers do report taxes on a bi-monthly basis, but just aggregate for the firm, no *individual* data is transferred.
4. That we all have an interest in accurate statistics leads us mostly to answer correctly, but, however real and important, this is not an immediate, direct motivation.
5. www.hr-xml.org.
6. Public register data could be used in order to save people time when typing data into such databases. The whole work history, the whole formal educational career could be captured automatically from the public databases – also guaranteeing its authenticity and accuracy – and more details could be accessible by hypertext methods.
7. Goetz Rohwer and Hans-Peter Blossfeld give a very good overview of event history analysis compared to other methods like panel data studies in their "Techniques of Event History Modeling", 2001, Lawrence Erlbaum Associates.
8. Employment data are available on establishment level.
9. Recently the register data contain physical co-ordinates of establishments in the so-called "Street, address and building" register. This means that one can "reconstruct" regional (imitate old, simulate new) regional administrative borders.
10. This is one extreme example of the lack of understanding of having a well-designed system to keep track of "firm demography" which is very common – even in the Nordic countries.
11. One cannot see the difference between these two number series, i.e. when you see a nine digit "organisation" number – you do not know if it is an enterprise or an establishment number. You have to use a computer to check. I would have made two different series, for example by starting the series with different first digits.
12. The International Standard Classification of Occupations is a very badly constructed indicator, but that is from a content point of view. Technically, it is a good "compact" indicator.