Virtual Sprint: 28-29 July:

What is a "virtual sprint"?

The concept behind this "virtual sprint" is for different members of a team based in different locations to work together on a common problem, or set of problems as part of a highly focused effort. This is not entirely consistent with the 12 Principles of the Agile Manifesto which says that the best way of communicating is "face to face conversation". However, it may be that this event is not so much a "sprint", which normally lasts several weeks, but more like a "hack", which normally only spans from a few hours to a few days. So, it may be more correct to call this a "virtual hack". But this doesn't really matter. This is partly just an experiment in working together in a different way and so we will see what happens....

How will it work?

The general approach for this virtual sprint has been borrowed from several "hack days" organised within the ONS Big Data team. This has involved team members based in two locations working on a common problem. One person takes the lead in organising the hack and the research tasks have related to a specific project that they are working on. Some preparation is involved, for example identifying and assigning tasks, and preparing test data sets etc. These team hacks have lasted about 5 hours with a initial phone meeting to ensure that everyone in briefed on what to do and a closing meeting to present and discuss any findings. The experience at ONS has been very positive. They have not only produced some very useful ideas, but they also proved very good for team building, especially across sites.

This "virtual sprint" will be organised on similar lines. The main difference is that it will last two days instead of one. There will be three Webex calls at the following times:

28 July: 9:30 to 10:00 - CET (Final briefing)

28 July: 16:00 to 16:30 - CET (Checkpoint)

29 July: 15:30 to 16:30 - CET (Presentation and discussion of results)

There will be a greater number of locations, but hopefully this will not be a problem. The most important thing is to bring a "hacking mindset". The allocated task and suggested approaches are only meant to be a guide and it is perfectly valid to pursue other ideas. Finding innovative solutions is the whole point of this exercise.

What are the potential tasks?

There seems to be sufficient support for at least one of the tasks to be on methods for **identifying duplicates**. It is clear that this will be a key challenge when combining data from different portals, but in some cases within job portals and certainly for job search engines. This will require some knowledge of record linkage and machine learning methods.

Since areas of expertise and possibly access to suitable tools and enviroments may be problematic for some countries, it would be desirable to have another task. This could be for example a exercise in collecting job vacancy data, possibly through the use of simple **web scraping tools**.

Another possibility is to do something with **CEDEFOP** data. ONS has made an initial assessment of the LMI system we have been given access to and our conclusion is that we really need access to the

underlying data. However, this might still be a good topic because the data is easily accessible.

What further preparation is needed?

You will get more out of the hack of you do a bit of preparation first. You may even want to do a small "pre-hack".

1. Task identification and allocation:

Please let me know which of these task(s) you would like to tackle:

1. Methods for identifying duplicates
2. Webscraping experiment
3. Assessment of CEDEFOP data

2. Background reading

If you are interested in the Robert Breton has put together a useful set of links on methods for identifying duplicates, This may give you some ideas on things to try.

3. Access to environment and tools

ONS have made an intial assessment of the Sandbox environment. It has R and R/Hadoop which is accessible via HUE . Python is also available via ssh, However, we are more familiar with our own sandbox environment and so we have decided not to use it for the hack. However, if you do would like to get access to the Sandbox please let me know and I will try to arrange it. Otherwise, you are on your own... please give some thought to what kind of tools you are going to need for your chosen task.

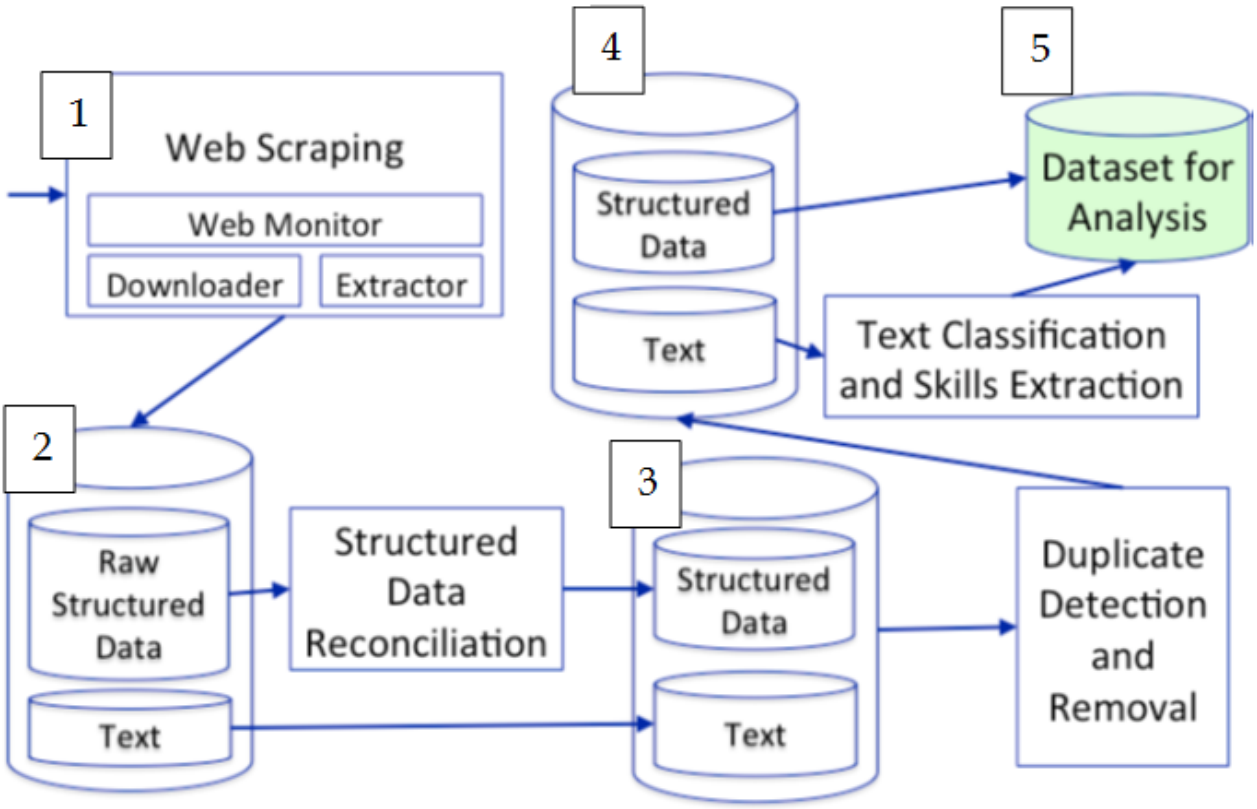
4. Access to Data

If you are interested in the duplicates problem, the ONS team are in the process of putting together a test dataset. This will be based several on UK job portals. We can make this available to anyone who needs some data, Alternatively, you can use your own data. Please let me know what you need

If you want access to the CEDEFOP LMI system, please send me an e-mail and I will provide the access details. This data is only to be used for ESSNet purposes.

Ideas from Dan Wu (Stats Sweden)

In the final report of project ”Real-time labour market information on skill requirements: feasibility study and working prototype”, a data model developed has been presented, see the figure below. This figure is used to explaining a general approach of web scraping for statistics. In Sweden, four data sets of job advertisements from the state employment agency have been exploited, concerning module 2, 3 and 5.



The data sets are xml files; they are then cleaned and transformed into a database, illustrated in module 2 to 3 in the figure. Since the data are from one source, duplication removal is not considered at the first place. Based on the data in module 3, we studied variables e.g. occupation, organization number and enterprise´s sectors.

The data sets have good coverage of occupations and the sectors. Only a small percentage of companies are covered comparing with the business register. However we cannot by the data itself conclude how good the coverage is. Data of other sources need to be complemented.

In the virtual sprint, the interesting questions can be for example:

1. How to map from the organization numbers in the adverts to the legal units in statistics of job vacancy, so that we can compare the advertisements data with the job vacancy statistics?
2. How to find the duplicate adverts in one source/multiple sources, can we combine the usage of the structured data and the text analysis?