Data Linkage Package

Lorraine Dearden (Institute for Fiscal Studies & Institute of Education)

2 July 2013

Introduction

- Two strands to this work
  - Facilitating linkage to Economic Data
  - Facilitating linkage to Education Data

- Economic Data is available across UK, but much more messy than a lot of data sets, particularly regarding the beginning and end of spells
- Education Data is generally of better quality, but collected differently in each of the four countries, therefore making consistent comparisons using this data across countries difficult
  - E.g. MCS research using education administrative data has so far only focussed on England (but now have Welsh data)
- Proposed work programme going to try and tackle these issues

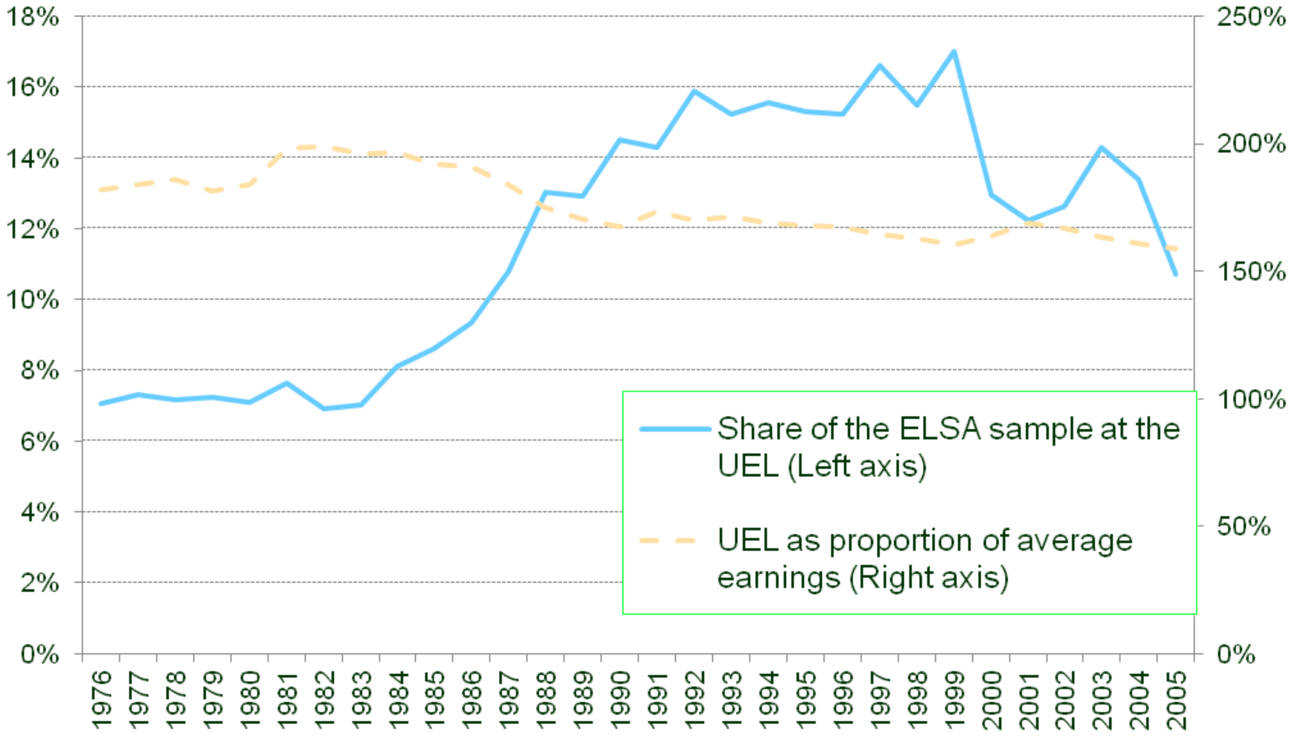
Potential of Economic Data Linkage Package

- The Government holds a number of key data sets that record individuals work, unemployment, and DWP benefit and programme history as well as earnings from PAYE, self-employment and National Insurance contributions.
- These data sets have the potential to give a detailed longitudinal picture of individuals economic activity going back as far as the 1940s (in the case of NI records)
- The economic data, however, is far from perfect
  - As we begin to link this data to the major longitudinal survey studies this will allow researchers to highlight more clearly the strengths and weaknesses of this data and how to mitigate well known problems with the data (e.g. employment spells ending on the 5th April).

National Insurance Data

- Government has collected NI contributions data back to the 1940s and holds this electronically from the 1970s
  - Means that we know how much a person earned if their salary was above the LEL and less than the UEL
  - This data has already been linked to ELSA and IFS has expertise in coding up this data to provide longitudinal data on earnings
  - From 1997 pay NI on earnings above the UEL so means that right censoring no longer a problem
  - Still problem for those below the LEL
- Coding up this data requires a detailed knowledge of the NI system from the 1940s
  - IFS has done this for ELSA and will be directly involved CLOSER work which will do it for other planned NI data linkage (NCDS, MCS)

UEL problem with NI data?



Note: Sample is ELSA wave 1 sample members who were successfully linked to their NI records. These people were aged approximately 24+ in 1976 and 52+ in 2004.

HMRC PAYE, Self-employment and Tax Credit data

- This data is available electronically from 1998 – records PAYE data from P14, P45 and P60 employer returns, and SE earnings from Self Assessment returns
  - Use just to record employment from 1998 – earnings only available from 2004
  - Also has data on maternity leave payments and statutory sick pay payments from 2004
- Tax Credit data is also available
- Data is currently being linked to ELSA and NCDS and MCS have put in request to link to this data (for those who consented in surveys for economic data linkage)
- This data has been linked to ILR data, DWP data and Student Loan Company data and other administrative data?

DWP Programme and Benefit Data

- Available from 1998 and records all spells of benefits and programme participation since April 1998
- There are differences in the extent to which background characteristics are recorded depending on programme/benefit
  - Linkage with NPD data would solve a lot of these problems but this hasn't occurred to date
- Data has been linked to HMRC PAYE data (both employment and earnings)– throws up lots of inconsistencies between two data sets (WPLS – Work and Pensions Longitudinal Survey)
  - Recently added Housing Benefit and Council Tax Benefit data for all people who have been a DWP client
  - But having both together allows one to apply sensible rules which can overcome some of the major problems with the data such as job spells starting on 6th April and finishing on 5th April – common in HMRC data
  - Linkage to Self Assessment and NI data would further enhance this but hasn't happened to my knowledge

What should we do with this data?

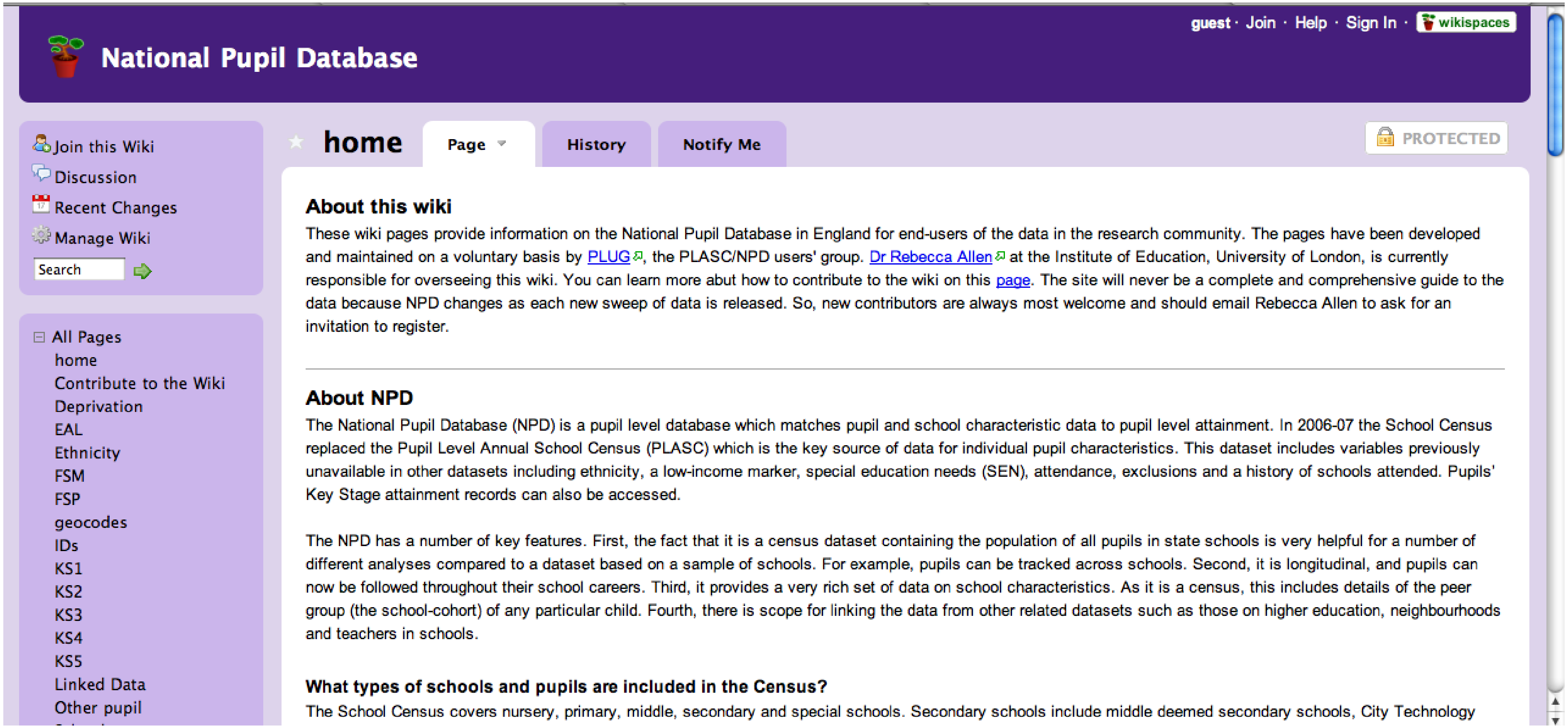
- IFS has developed quite a bit of knowledge around this data with its ELSA work
- The proposed CLOSER work which should further enhance our knowledge and enhance proposed economic data linkage to all of the major longitudinal surveys running in the UK
- Potential of this data is huge – particularly the NI data which goes back to the 1940s
- Need to think carefully about derived variables from this data set
  - For a large majority of researchers summary variables may be sufficient and it may be possible to have access to these variables via normal access routes
  - Other forms of the data will need more secure access – but this will need to be negotiated with data owners

Potential of the Education Linkage Package

- Attempt to create a consistent (and hopefully more UK focussed) education administrative data set which is as consistent as possible over time and across countries
  - Have been deliberately ambitious but feel confident that we can get someway to achieving these ambitions
- Draw on our existing expertise in using a wide range of education administrative data
- We already know a lot about the strengths and weaknesses of the data and common mistakes made in coding the data
  - This knowledge is already shared and the CLOSER package will help cement this knowledge transfer further

Collaborative ‘help’ pages

http://nationalpupildatabase.wikispaces.com



About this wiki

These wiki pages provide information on the National Pupil Database in England for end-users of the data in the research community. The pages have been developed and maintained on a voluntary basis by PLUG, the PLASC/NPD users' group. Dr Rebecca Allen at the Institute of Education, University of London, is currently responsible for overseeing this wiki. You can learn more abut how to contribute to the wiki on this page. The site will never be a complete and comprehensive guide to the data because NPD changes as each new sweep of data is released. So, new contributors are always most welcome and should email Rebecca Allen to ask for an invitation to register.

About NPD

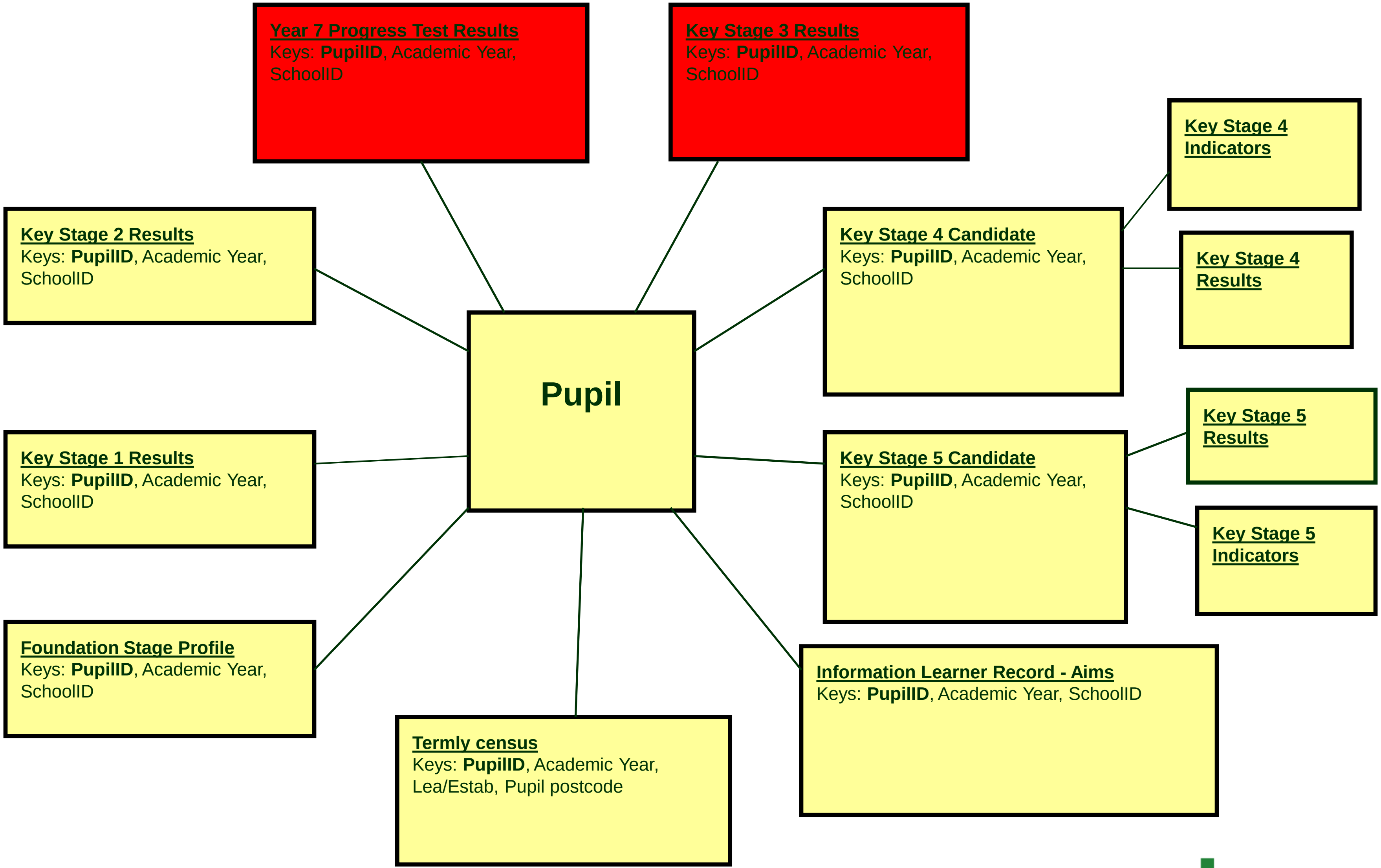
The National Pupil Database (NPD) is a pupil level database which matches pupil and school characteristic data to pupil level attainment. In 2006-07 the School Census replaced the Pupil Level Annual School Census (PLASC) which is the key source of data for individual pupil characteristics. This dataset includes variables previously unavailable in other datasets including ethnicity, a low-income marker, special education needs (SEN), attendance, exclusions and a history of schools attended. Pupils' Key Stage attainment records can also be accessed.

The NPD has a number of key features. First, the fact that it is a census dataset containing the population of all pupils in state schools is very helpful for a number of different analyses compared to a dataset based on a sample of schools. For example, pupils can be tracked across schools. Second, it is longitudinal, and pupils can now be followed throughout their school careers. Third, it provides a very rich set of data on school characteristics. As it is a census, this includes details of the peer group (the school-cohort) of any particular child. Fourth, there is scope for linking the data from other related datasets such as those on higher education, neighbourhoods and teachers in schools.

What types of schools and pupils are included in the Census?

The School Census covers nursery, primary, middle, secondary and special schools. Secondary schools include middle deemed secondary schools, City Technology

What data is included within English NPD?



Pupil

Year 7 Progress Test Results
Keys: **PupilID**, Academic Year, SchoolID

Key Stage 3 Results
Keys: **PupilID**, Academic Year, SchoolID

Key Stage 2 Results
Keys: **PupilID**, Academic Year, SchoolID

Key Stage 1 Results
Keys: **PupilID**, Academic Year, SchoolID

Foundation Stage Profile
Keys: **PupilID**, Academic Year, SchoolID

Termly census
Keys: **PupilID**, Academic Year, Lea/Estab, Pupil postcode

Key Stage 4 Candidate
Keys: **PupilID**, Academic Year, SchoolID

Key Stage 4 Indicators

Key Stage 4 Results

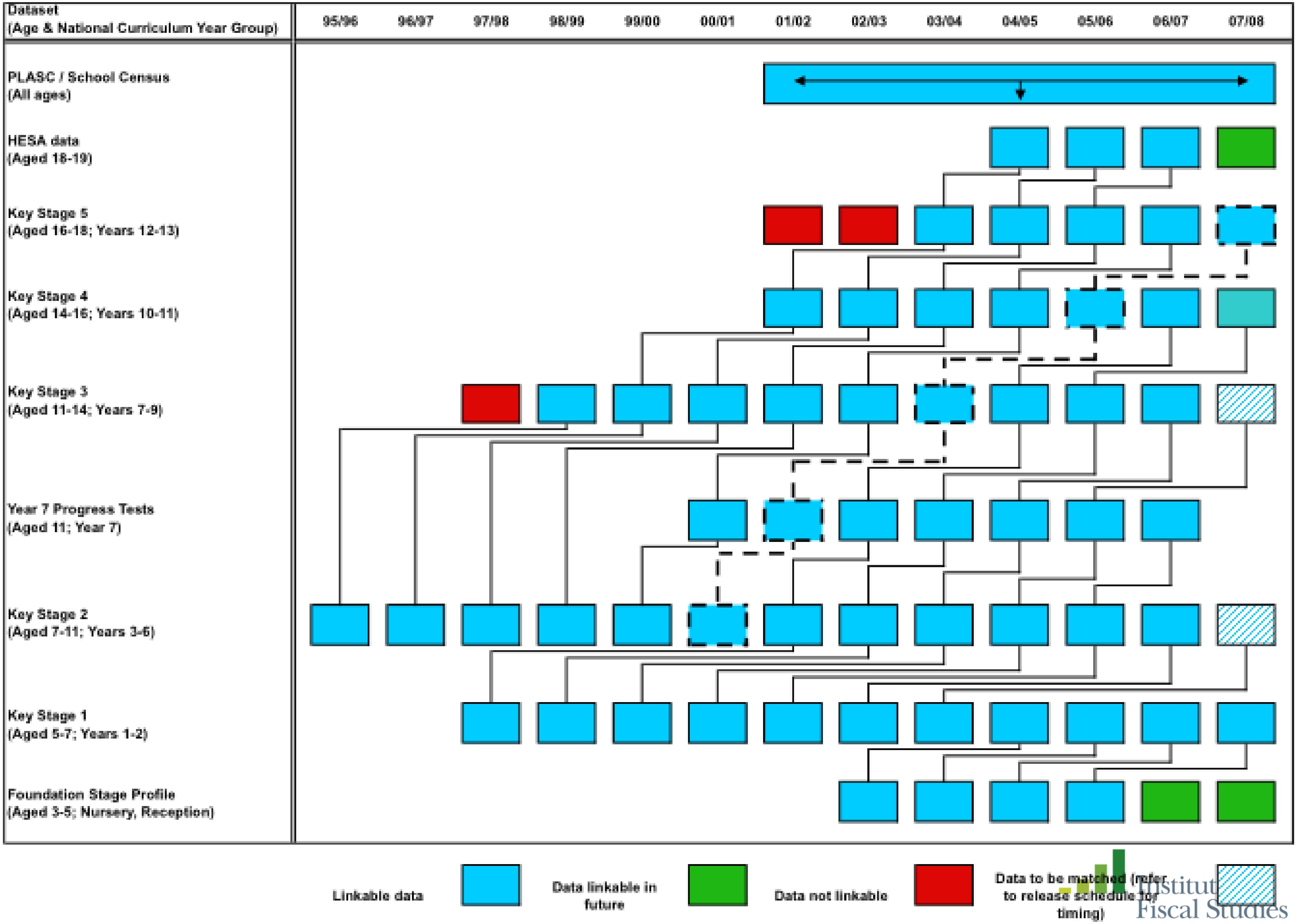
Key Stage 5 Candidate
Keys: **PupilID**, Academic Year, SchoolID

Key Stage 5 Results

Key Stage 5 Indicators

Information Learner Record - Aims
Keys: **PupilID**, Academic Year, SchoolID

Pupil tracking from FSP to HE in the National Pupil Database



Dataset (Age & National Curriculum Year Group)

95/96 96/97 97/98 98/99 99/00 00/01 01/02 02/03 03/04 04/05 05/06 06/07 07/08

PLASC / School Census (All ages)

HESA data (Aged 18-19)

Key Stage 5 (Aged 16-18; Years 12-13)

Key Stage 4 (Aged 14-16; Years 10-11)

Key Stage 3 (Aged 11-14; Years 7-9)

Year 7 Progress Tests (Aged 11; Year 7)

Key Stage 2 (Aged 7-11; Years 3-6)

Key Stage 1 (Aged 5-7; Years 1-2)

Foundation Stage Profile (Aged 3-5; Nursery, Reception)

Linkable data 🟦 Data linkable in future 🟩 Data not linkable 🟥 Data to be matched (refer to release schedule for timing) ▨

Existing linkages to surveys

- ALSPAC
- Millennium Cohort Study
- Longitudinal Study of Young People in England
- Labour Force Survey (age 19-21 respondents from 2011)
- Understanding Society

Existing linkages to other education admin data

- HEFCE data
  - seven cohorts of children have been linked to HEFCE data (the first being those that sat KS4 in 2001/02 and KS5 in 2003/04 who have been linked to HEFCE data from 2004/05.
  - The HEFCE data records HEI, course, drop-out and eventually results (class of degree).
  - HEFCE data has been linked to UCAS data but the three data sets have not been linked into one (UCAS very sensitive about their data)
- NISVQ/ILR Data
  - Vocational courses and training
  - This data has been linked to LFS but not to NPD *and* LFS together
  - This data has been linked to HMRC/DWP data but not also to NPD data

What should be CLOSER priorities

- Careful balance between creating broadly consistent data across countries vs using all available detail in particular countries, particularly English
- Lots of inconsistencies between old and new NPD data, school identifiers often tricky because of LA changes and school governance changes
  - Many researchers completely unaware of this
- Other issues.....