ADVISORY PANEL ON CONSUMER PRICES – STAKEHOLDER

Consumer Prices – Alternative Data Sources Roadmap

Status: (final)
Expected publication: (alongside minutes)

Purpose

1. The Consumer Prices Data Collection Strategy presented in Paper APCP-S(18)03 set out a move from collecting prices manually to electronic means wherever feasible and efficient. Work is required in four key areas – obtaining robust sources of data, development of systems, methods research and finally assessing the impact on headline consumer price statistics.
2. The paper in Annex A sets out the work achieved so far in these areas and presents a more detailed roadmap for the implementation of these alternative data sources. This includes milestones for when key research papers will be brought to both advisory panels for review.

Actions

3. Members of the Stakeholder Panel are invited to:
   a) provide feedback on the research to date
   b) comment on the timelines and scope of the alternative data sources project

Tanya Flower
Prices Division, ONS
January 2019

List of Annexes

Annex A	Consumer Prices – Alternative Data Sources Roadmap

Annex A – Consumer Prices – Alternative Data Sources Roadmap

1. Introduction

Paper APCP-S(18)03 presented to APCP-S in February 2018 set out a high-level vision for the Consumer Prices Data Collection Strategy, focusing on a move from collecting prices manually, either on the high street or by trawling websites, to electronic means wherever feasible and efficient. It included a high-level work plan setting out four key areas of further work required – obtaining robust sources of alternative data, methods research, assessing the impact on consumer price statistics and the development of systems to support the inclusion of new data sources.

This paper sets out the work achieved so far in these areas and presents a more detailed roadmap for implementation of these alternative data sources. This transformation will be the largest change to consumer price statistics in a generation, and the scale and importance of this work should not be underestimated. We will be reliant on developments in many areas, including new technology platforms.

There will also be a need to research and develop new methods for processing these data for the purpose of calculating consumer price indices. We introduce these research areas alongside projected timeframes for completion and recommendations. Due to the scope of the work required, we propose that this workplan will be separated out from the main consumer prices development plan in future. This means that the alternative data sources workplan will be discussed and prioritised separately at the May 2019 APCP meetings and then published alongside the main development plan when the annual report is published next year.

This roadmap presents our stakeholders with a detailed timeframe for the implementation of these new methods and data, and signposts the opportunities along the way to provide comments and feedback on the new developments we will be proposing. Our ambitious plan for implementation is to include alternative data sources in the production of our aggregate measures of consumer price statistics by January 2023, although there will also be experimental indices produced at various intervals before the final implementation date.

2. Achievements over the last year

In 2018, the focus was on obtaining robust alternative data sources and completing exploratory analysis on suitable methods that can be used to process these data from the raw input files to final item level indices. The analysis was carried out using the test part of our new strategic IT platform (DAP E), which has enabled us to process bigger and ever more detailed datasets.

By the end of 2018, the status of the 2 possible alternative data sources (web scraped and scanner data) are as follows –

1. Web scraping – we are now receiving web scraped data from mySupermarket. These data are for around 25 categories in the basket covering areas such as clothing, electronic items and package holidays (for a full list, please see Table 1). There are no back series with these data so we will need to build up a sufficient time series before a final impact assessment can be completed. Work has begun on checking to see if the Robot Tool[1] will be useful for production in the short term. We are also exploring a web scraped data source for used cars.

[1] The robot tool is essentially a price monitoring tool that notifies collectors when there has been a change in price of an item advertised online. It is suited to items that do not change price frequently (for instance, 'golf club membership fees'). It is not worth setting up a web scraper for these items because of the infrequent price changes, but the robot tool can monitor these items and notify collectors the day that the price changes.

2. Scanner data – we are continuing to engage with retailers on receiving some transaction data, targeting some of the largest retailers that we currently collect prices from. This will cover areas of the basket such as groceries and clothing. Towards the end of 2018, we received a test dataset from one of these retailers and we are aiming to receive some regular data feeds for several retailers by April 2019. We may be able to receive a historical back series with scanner data, so less time may be required for the impact analysis. We will also have access to a database of rail fares transactions from February 2019.

Table 1: List of categories covered by mySupermarket

One-item datasets	**Multi-item datasets**
Blu-rays	Airfares
Carpets*	Books*
CDs	Clothing*
Desktops	Computer accessories*
DVDs	Jewellery*
Laminate*	Package holidays
Laptops	Personal articles*
Luggage*	Sporting goods*
Printers	Sports food supplements*
Routers	Stationary*
Rugs*	Shoes*
Smart phones	
Tablets	

Note: This table has been separated into areas where we have one dataset per item (as defined in the ONS classification structure), for example blu-rays, and categories where there is more than one item per dataset, for example clothing. In total there are about 150 representative items that are being collected by mySupermarket. * indicates categories where a large proportion of price quotes are collected locally in the current collection.

In terms of exploratory methods analysis, Paper APCP-T(18)13 presented to APCP-T in September 2018 summarised the work carried out so far and proposed a high-level end to end pipeline, comprising of individual modules required to process the data, for example “classification”.

By the end of 2018, we were able to develop a prototype version of this pipeline on the new test version of our future strategic processing platform (DAP E). This pipeline took mySupermarket web scraped data for laptops, cleaned and processed it, and produced some final output item level indices using a number of different elementary aggregate formula. Although there will not be much of a back series, we are aiming to publish these indices in April 2019 with some initial discussion about how we have calculated them as part of our ongoing publication strategy (discussed in detail in the next section).

3. Roadmap to 2023

In this section, we present a high-level roadmap that summarises the key phases of our implementation plan. Further detail around research and publications are also provided. During this period, we will also be liaising regularly with our users and the Office for Statistics Regulation to ensure that users have the opportunity to feed into our planning, with the intention being to have a formal consultation period in 2022 before the final implementation goes live.

The implementation plan is based on a staged approach to incorporating alternative data sources for different areas of the COICOP basket. This allows us to better isolate the impact of changing data sources and methods for particular categories. Changes to the first tranche will be incorporated in January 2023, and then further categories will be incorporated on a rolling annual basis beyond this point.

We may also consider a staged approach to introducing these new data and index methods. In the first phase, we could draw a sample from these alternative data sources and apply a fixed base method as we currently do (essentially replicating the current methodology but with a new underlying data source). At the next stage of implementation, we could then move to more sophisticated methods (for example, multilateral methods), which will make better use of the greater coverage of these alternative data sources.

In the first phase of the project, we will focus on categories where we either have or are confident in receiving data for shortly. These include:

1. Technological goods (for example, laptops)
2. Chart collected items (CDs, DVDs, Blu-Rays and books)
3. Package holidays
4. Clothing
5. Rail fares
6. Used cars
7. Groceries (depending on which retailers we are able to secure scanner data from)

Throughout the rest of this section, any reference to experimental item indices or selected COICOP groups refer to those identified in the list above.

Figure 1 presents our high-level milestones for incorporating alternative data sources in our consumer price statistics, with further detail on research, publications, systems and other workstreams presented in Figures 2 and 3.

Figure 1: High-level milestones for incorporating alternative data sources in our consumer price statistics

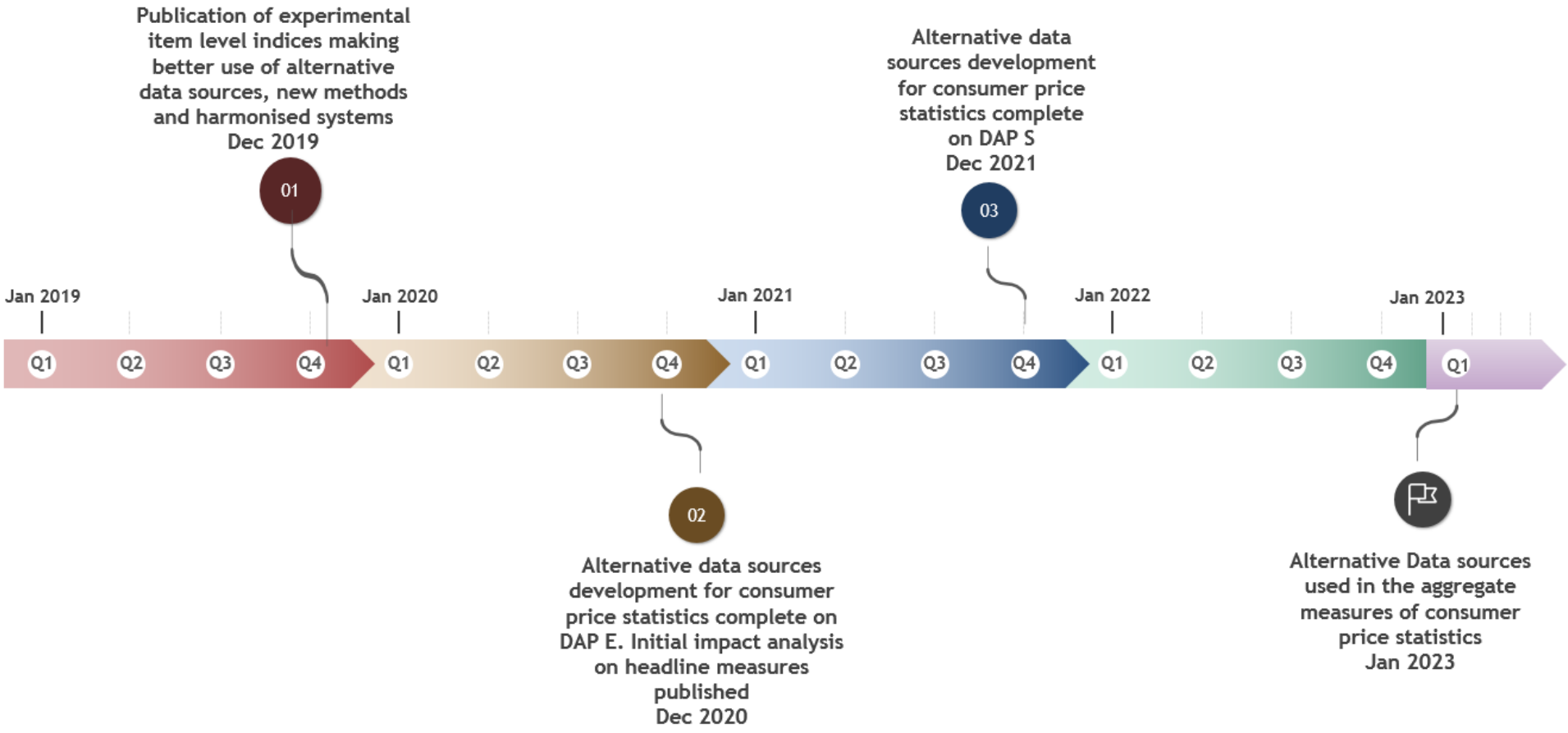


Figure 2: Research and publication milestones

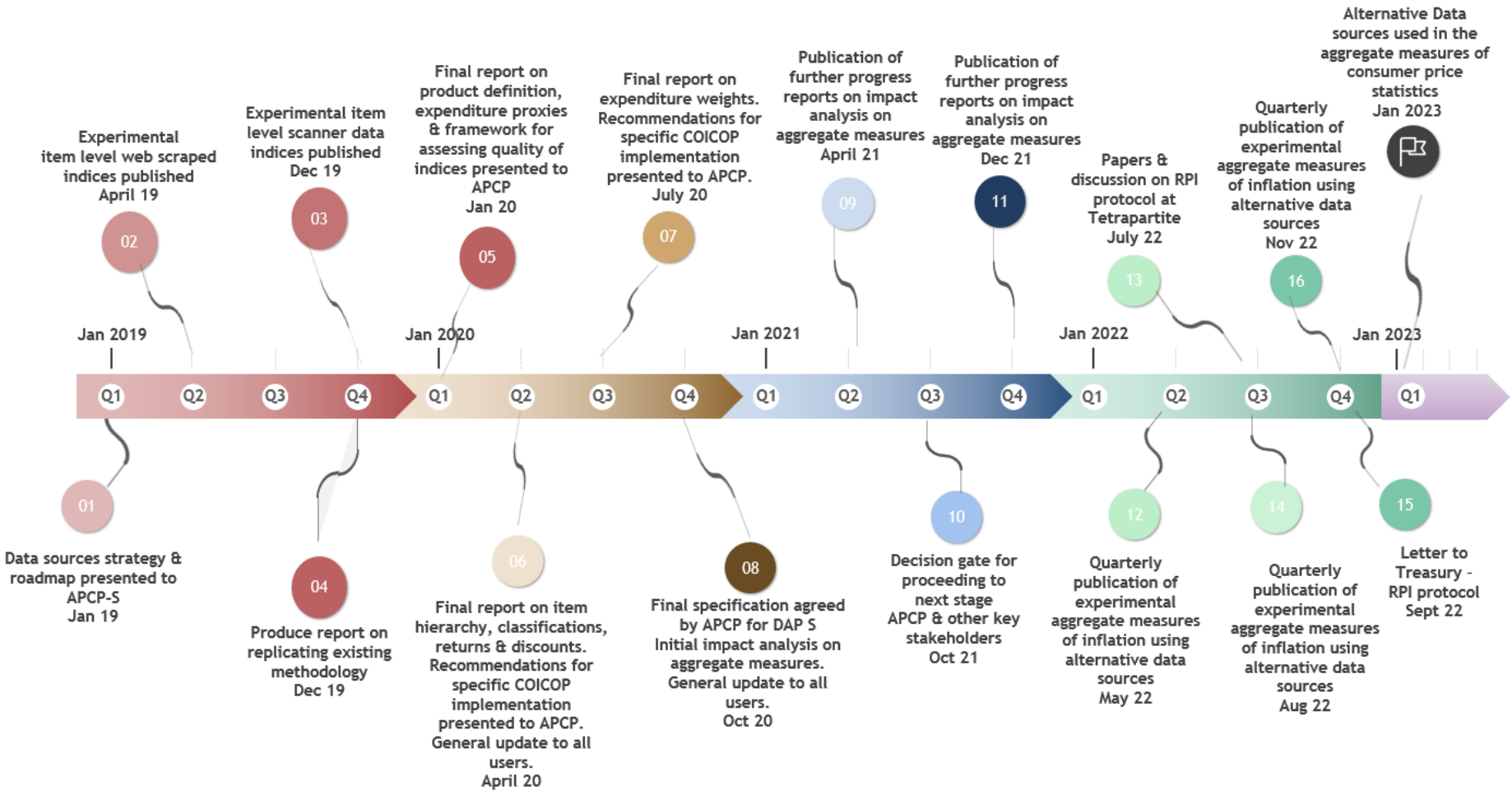
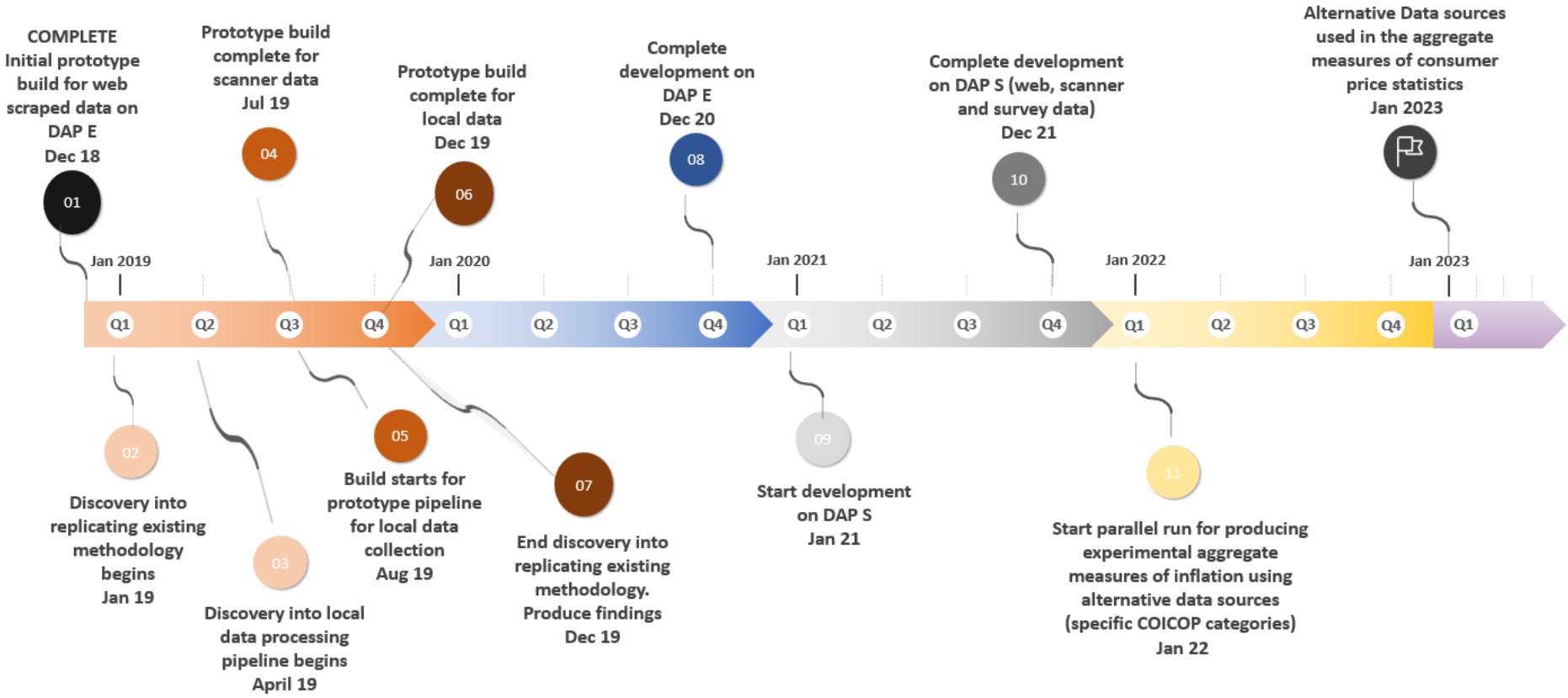


Figure 3: Discoveries & system developments



The first high-level milestone in December 2019 includes the publication of experimental item level indices calculated from web scraped data (initial findings presented in April 2019 – Figure 2) and scanner data (December 2019). This expands the work completed by end of 2018, where a prototype pipeline for constructing item level indices for laptops data was constructed on the test part of our new strategic IT platform (DAP E). The framework used to build this pipeline (i.e. the different modules), are designed to be applicable to all items and data sources that may be used to construct consumer price statistics.

The end of 2019 will also see a report published on whether or not the new data sources can be used as a direct replacement for our existing sample of prices (Figure 2). This would also mean replicating the existing method of choosing comparable and non-comparable replacements. This is not the best solution in the long-term as it does not make full use of the benefits these new data sources can bring in terms of increased coverage and higher frequency collection. However, it may be helpful in a staged approach if we wanted to replicate the existing methodology in the first phase before moving onto more complex methodology.

Work will also begin on harmonising our systems. In 2019, a discovery will start on lifting the current production system used to process the locally collected data onto the new DAP E platform. By the end of 2019, the DAP E laptops pipeline will have also been expanded to include processing scanner data and locally collected data for selected COICOP categories.

There are a number of research projects that will begin in 2019 and produce final recommendations towards the beginning and mid-2020 (Figure 2). These include:

1. Framework for assessing the quality of consumer prices indices produced using alternative data sources. This work will summarise the properties of a desirable index calculated using these “big” datasets and provide recommendations on how a final index method and staged implementation plan could be selected for different COICOP groups. The recommendations from this will feed into our final decision on which method/s to choose for implementation.

2. Expenditure weights for web scraped data. One of the limitations of web scraped data is that it doesn’t provide information on expenditure. This work will identify if the lack of expenditure weights introduces any bias into any index based on web scraped data, and if we can approximate expenditure weights using alternative data sources like page rankings.

3. Reviewing the existing item hierarchy and definitions. Our current methodology is based on representative items, which are generally quite narrowly defined. This work will review the existing product hierarchy considering the increased coverage of products in alternative data sources.

4. Classification techniques. New methods will be required to process these new “big” data sources, including new ways of automatically classifying products to a specific COICOP category. This work will recommend which methods are suitable for particular categories.

5. Product definition (item vs groups). In the current methodology, an individual product is followed over time and compared back to the base period. An alternative approach would be to follow the average price of a defined group of homogenous products instead. International research has shown this to be a viable alternative for categories such as clothing which experience high rates of product churn over time.

6. Expenditure weights for different data sources/retailers. This work will recommend methods and suitable data sources that will allow us to aggregate together data sources

from different collection methods for the same category (for example, locally collected data for bread from local bakeries alongside scanner data from a large retailer).

7. The impact of product returns and discounts on alternative data sources. The issue of returns affecting expenditure weights for particular categories (for example, clothing) may impact on how we can use expenditure weights in a final item index.

Work on many of these projects have already begun (for example, a paper on expenditure weights for web scraped data was taken to recent APCP-T meetings in September 2018 and January 2019). There are also a number of national statistical institutes following similar programmes of work, and a number of study visits are planned throughout 2019 and 2020 to make best use of international experience for our analysis. It is envisaged that various iterations of papers will be produced either for presentation at APCP-T and APCP-S, or sent round for feedback via correspondence, with the final recommendations being presented at various stages during 2020.

Depending on the results from these projects, the final part of 2020 will also see recommendations made on the specific implementation of alternative data sources for the particular COICOP categories listed above. This will also include an initial impact assessment carried out on how these changes will affect the aggregate consumer price statistics measures (the second high-level milestone in Figure 2).

These projects will all be carried out and delivered using the DAP E platform and the associated pipeline. DAP E is a test environment that provides a useful research tool to explore and develop new methods of processing large datasets. The final production system will be built on DAP S, a second platform which is designed to support the regular production of national statistics. The recommendations that will be delivered by the research projects above on DAP E will be translated into programming requirements for DAP S by the end of 2020.

During 2021, we will continue to produce experimental indices and the resulting impact analysis for users to monitor. This is in combination with building the final processing pipeline on DAP S, which will be completed by December 2021 (the third high-level milestone in Figure 1).

Finally, in 2022 we will run a full parallel year of processing these data sources and outputting experimental versions of the aggregate statistics on a quarterly basis. Any changes required to the RPI will need to be discussed with HMT and the Bank, with the letter summarising any impact sent in September 2022. The final implementation date for the first tranche of COICOP categories will be January 2023.

While the final implementation date of 2023 is the first time the data will be used in production, there may be areas where the data can be used more quickly to benefit the existing collection. For example, the collection of price and characteristics data for our current hedonic models is currently done manually online. While not replacing the specific products followed over time by the index, we could replace the data used to construct these models using web scraped data in the short-term. There are also options to use these data to add an additional layer of quality assurance for our existing collections.