Report of a breach of the Code of Practice for Statistics

Core Information

Title and link to statistical output	Unistats dataset (2020/21)
Name of producer organisation	Higher Education Statistics Agency (HESA)
Name and contact details of person dealing with report	Rebecca Mantle rebecca.mantle@hesa.ac.uk
Name and contact details of Head of Profession for Statistics or Lead Official	Jonathan Waller jonathan.waller@hesa.ac.uk
Link to published statement about the breach (if relevant)	
Date of breach report	02 November 2020

Circumstances of breach

Relevant principle(s) and practice(s)	Practice T3.6: Statistics should be released to all users at 9.30am on a weekday.
Date of occurrence of breach	21 October 2020

Give an account of what happened including roles of persons involved, dates, times etc.

The Unistats dataset contains information about courses offered by higher education providers. It is published by HESA on behalf of the Office for Students. The 2020/21 dataset was originally scheduled to be published at 09:30 on 28 September 2020, but delays were pre-announced through HESA's publication schedule following the identification of a number of quality issues that were taking longer than expected to resolve. A revised publication date of 21 October 2020 was pre-announced alongside a full explanation for the delay.

At 09:20 on 21 October, HESA's Official Statistics Team were alerted to a further quality issue in the dataset and it was determined that this was a data error that needed further investigating and correcting prior to publication.

The issue discovered related to a small number of missing courses from the dataset. While the data for these courses had been collected, in producing the output these were deemed initially to be duplicate courses and were excluded from the dataset. Usual Quality Assurance (QA) activities did not pick up on this issue. Additional more rigorous QA work which would not have usually been part of the process was undertaken in the week leading up to publication to provide further reassurance on the quality of the output.

While this activity initially did not pick up any new issues, this particular error was identified the evening prior to release by the production team. The team worked at maximum pace to investigate and resolve this issue in time for the publication so as to avoid a further delay to the release. However, it became apparent the following morning that time was running too short to implement and sufficiently test a resolution. The release was not made at 09:30 and users were informed as soon as practically possible, but shortly after 09:30 (at 09:35) thereby breaching the Code of Practice for Statistics.

When the extent of the error became clear and it was possible to state with confidence when the dataset would be updated and sufficiently quality assured, a new revised publication date of 28 October 2020 was announced and met on time.

The cause of the breach lay largely in issues with the process to produce and quality assure the dataset. The dataset itself is extremely complex in structure and content, as are the associated release processes.

The current approach does not provide an adequate platform to fully test and challenge specification changes prior to implementation and coding. In addition, when quality issues are identified during production the skills and knowledge to resolve them are concentrated in too few staff, resulting in lack of resilience and risks of publication delays.

Impact of the breach

Provide details of the impact of the breach both inside the producer body and externally

Any users waiting for the 2020/21 dataset at 09:30 on 21 October would have noticed that the webpage had not been updated and information contained within it still referred to the 2019/20 dataset. At 09:35 an announcement on that webpage was made, in addition to our upcoming data releases webpage to clarify that the publication would be delayed further due to continuing essential work to rectify quality issues. We are not aware of any adverse public reaction to the news regarding the further delay.

There was an impact on the update to the Discover Uni website which presents the Unistats dataset in a user friendly and interactive format. The 2020/21 dataset was due to be used by the Office for Students (OfS) to update the Discover Uni website shortly after the release of the Unistats dataset. Due to the further delay and following communications with the OfS the Discover Uni website was updated on 5 November 2020.

Corrective actions (taken or planned) to prevent re-occurrence

Describe the short-term actions made to redress the situation and the longer term changes to procedures

As soon as it became apparent that the data error meant the dataset was not of sufficient quality to publish, an announcement was made both on HESA's upcoming data releases webpage and also on the webpage for the dataset itself. A new publication date was not announced until some days later when we could provide full confidence of being able to meet it. A conversation also took place with the OfS to alert them of the situation and discuss the impact and revised scheduling of the update to the Discover Uni website.

In relation to future corrective actions, as we have had to submit three previous breach reports related to this dataset we are holding a 'lessons learnt' session to review the current production and quality assurance processes.

We expect a fundamental overhaul of the process that would see HESA's Official Statistics Team leading work right from the start of the process to review and develop the data specification and also create a new process to produce a version of the dataset in parallel and independently from the developers who currently produce the dataset.

This would strengthen quality assurance, with a more robust process for identifying ambiguities or edge-cases in the data specification and to provide a means of verifying the production data set. In addition to this, we are working with the Good Practice Team to learn more about Reproducible Analytical Pipelines and how they may be able to help simplify our processes.