The Pros and Cons of Automatic Data Extraction -

A cautionary tale from Ireland about modernising the collection of Earning Statistics

Moore, Ken
Statistician, Short Term Statistics
Central Statistics Office,
Skehard Road,
Ireland
ken.moore@cso.ie

&

MacFeely, Steve
Director of Business Statistics & Innovation,
Central Statistics Office,
Skehard Road,
Ireland
steve.macfeely@cso.ie

Abstract:

In 2008 the Central Statistics Office in Ireland introduced, with the cooperation of all the major payroll (software) systems providers, a facility whereby statistical returns for the *Earnings, Hours and Employment Costs Survey* (EHECS) could be returned automatically via XML.

In 2010 the ambitions and design of this survey were reassessed following a review to understand why the original objectives were not being achieved. From this review a number of important lessons were learned. This paper presents a frank summary of the strengths and weaknesses (pros and cons) of this approach to data collection.

Keywords:

Earnings, XML, response burden

Introduction

In 2008 the Central Statistics Office in Ireland introduced a new quarterly earnings survey; the *Earnings, Hours and Employment Costs Survey*. This new survey replaced a suite of basic sector specific surveys with a significantly more comprehensive survey common to all economic sectors. Given the very comprehensive nature of this new questionnaire (and the increased response burden it would impose), the CSO in cooperation of all the major payroll (software) systems providers operating in Ireland developed a facility whereby completed questionnaires could be returned automatically via XML. In 2010 the ambitions and design of the EHECS were reassessed to understand why the original objectives had not been achieved.

This paper is divided into five sections. Section 1 provides some background and details the original ambitions of the EHECS survey. Section 2 outlines the initial reality and apparent failure of the system introduced. Section 3 details the review undertaken in 2010, Section 4 lists the benefits arising once the system had been tweaked and Section 5 concludes with some of the pros and cons of automatic data extraction together with a summary of the lessons learned to date.

1. Original aims and objectives

Prior to the *Earnings, Hours and Employment Costs Survey*, the collection of earnings data within CSO had been fragmented and inconsistent. The data available differed across the economic sectors and was published on different schedules, frustrating users. One of the primary aims of EHECS was to provide a harmonised and consistent dataset across all economic sectors.

Another ambition of the EHECS project was to provide more comprehensive information on employment and earnings. It was recognised that these additional data requirements would result in increased respondent burden. To try and compensate for or counteract against this, the CSO in cooperation with the main 29 payroll providers operating in Ireland developed a facility to extract the required information directly from both their payroll and their time-attendance systems[1]. It was reasoned that as these data were readily available from payroll/time-attendance systems, XML returns could be made both easily and promptly, thereby transmitting high quality data in a timely manner with little or no intervention required from the respondents.

Thus the overall aim of EHECS was to provide a "win-win" solution whereby the quality and detail of the earnings data would be improved, the costs of compiling the data would be reduced and respondent burden would also be reduced through the provision of a modern "push button" data collection system.

[1] Technical specifications were also made available to those enterprises that used bespoke systems.

2. Failure

By 2009 it was clear that the aims and objectives set out for EHECS were not being achieved. Across a broad range of metrics, from timeliness to response burden, problems were apparent. The burden in particular was of concern for two reasons:

1. EHECS appeared to generate more burden than anticipated with respondents complaining that the survey was too long, too detailed and too complex. Industry representative groups also accused CSO of "gold plating" as the requirements were in excess of that specified under EU legislation. In hindsight the questionnaire more resembled a detailed annual survey than a short-term inquiry form. The size and complexity of the questionnaire was reflected in the completion times, which on average were 108 minutes. An added complexity arose from the fact that the information sought required the involvement of both Payroll and HR functions within an enterprise. So while the payroll elements of the questionnaire were easily populated, many of the non-payroll data were not readily available. This resulted in significant time delays for CSO and considerable frustration for the responding enterprises.
2. The problems created by EHECS appeared to ripple out, creating an adverse affect on other CSO enterprise surveys, several of whom reported increased levels of dissatisfaction regarding response burden among responding enterprises.

Timeliness also quickly emerged as a cause for concern. One of the primary objectives had been to achieve compliance with the T+70 days deadline specified by EU legislation[2]. With an average delivery time of T+207 days for 2009, it was clear this deadline was far from being achieved.

One of the main challenges arising from collecting data via multiple payroll systems is scheduling and coordination. CSO must try and ensure that any amendments or updates of the electronic module are properly coordinated across the 29 payroll suppliers and numerous bespoke systems. Large multinational enterprises (particularly in the retail sector) pose an additional challenge as many of their headquarters are overseas and use more complex customised systems. This became a major issue when CSO implemented significant system changes to rectify the problems outlined above.

3. Facing up to reality

In light of the lack of success (resulting from the issues noted above), a top to bottom review of operations was conducted in early 2010. As part of this review, respondents, payroll software suppliers and staff were consulted. The review brought clarity by prioritizing a number of competing issues. Specifically it was agreed that the focus of the EHECS should be a short-term indicator and that the XML solution should be "push button" and any content that interfered with this should be removed.

[2] Council regulation (EC) No. 1165/98 of 19 May 1998 concerning short-term statistics (OJ L 162, 5.6.1998, p.1)

Sweeping changes were proposed and these were accepted by senior management. The changes, implemented from Q2 2010 are summarised below:

1. Different Survey forms were issued depending on size of the enterprise. Large enterprises (i.e. those with 100 or more employees) or enterprises who made returns electronically via XML received a detailed EHECS questionnaire. Small enterprises (i.e. those with less than 100 employees) received a shortened, less detailed questionnaire;
2. The content of *all* survey forms was greatly reduced. Non-payroll related questions were removed where possible (e.g. "*paid hours not worked*" was no longer required);
3. The requirement to provide details on occupation split was removed from the questionnaire sent to small enterprises; and
4. In recognition of the initial set-up effort required of an enterprise to return data via XML, a specialist support team was established within CSO to assist enterprises through this phase.

The loss of data in terms of both EU requirements and national use arising from these changes was low. Occupational breakdown were lost for enterprises with less than 100 employees but with XML returns retaining this detail, good quality estimates could still be imputed. The only significant data loss from an EU perspective would be the loss of the *paid hours not worked* variable on a quarterly basis and would now be imputed to derive hourly earnings for the quarterly Labour Cost Index.

4. Success at last

The changes made to the EHECS survey, systems and procedures appear to have worked sufficiently that the project is now considered a success. Overall, total and partial non-response has fallen significantly, as has the volume of editing and imputation required. A number of standard management metrics are presented in this section to illustrate and quantify this success; timeliness; burden reduction; resources consumed; response rates; number of edits required; and share of XML returns (see Appendix 1).

4.1 Timeliness

Timeliness is one of the key components of quality. During 2008 and 2009 the delays in publication were significant, with provisional data being issued some 38 weeks after the reference period and final taking over a year (55 weeks). Since the second half of 2010 the time lags have stabilised around 10 weeks for provisional data and 23 weeks for final data. While further improvements in timeliness will be harder to achieve, the long term aim is to achieve T+8 and T+15 weeks. A very positive sign for the future is now that some enterprises make their returns before the survey forms have issued.

4.2 Burden Reduction

The introduction of the EHECS questionnaire imposed a serious burden to business; with a sample size of 10,500 enterprises per quarter (41,500 enterprises for the year 2008) and an average completion time of almost 108 minutes, the total burden (adjusted for non response) was 42,700 hours or the equivalent of a total administrative cost of €1.9 million (in 2005 prices). By 2010 (the latest published burden metrics available) the sample had been more reduced to less than 7,500 per quarter (29,300 enterprises for the year 2010), the average completion time had been reduced to 63 minutes and total

administrative costs to business had fallen to €833,000. It should be noted that for enterprises availing of and using the XML facility, the completion time is now less than 5 minutes per form.

4.3 Resources

A key concern of the EHECS project during 2008 and 2009 was the number of staff working in the area. At the beginning of 2010 there were 40 FTE staff (across all grades) dedicated to EHECS. By May 2011 the compliment of staff had been reduced by 26% to 29.5 staff. It is planned to release a further 4.5 staff in the autumn of 2011, yielding an aggregate reduction of 37%.

4.4 Response Rates

Another area of concern in during 2008 and 2009 was the response rate. At the beginning of 2009 the response rate was 45% which only represented 49% of total employment. At the end of the first quarter in 2011, response rates had improved to 57% which improved the coverage of total employment to 71%. The longer term aim is to improve enterprise response rates to at least 65%.

4.5 Edits and partial non-response

Micro-editing is expensive and may not necessarily improve the aggregated data. Furthermore too many edits suggests there is something fundamentally wrong with the data collection. Prior to the changes made in Q2 2010, over 35% of returns failed at least one edit check. The simplification of the questionnaire dramatically reduced this to just over 20%. Equally partial non-response fell from 23% to 15%.

4.6 Share of XML returns

The CSO has invested very considerable time and resources into designing the XML module for EHECS. The more enterprises that avail of this facility the more the quality of the data improves and response burden falls. Overall, electronic returns to CSO have remained largely unchanged at between 30 – 35%. However the composition of these electronic returns has changed considerably. The contribution of returns made via excel has fallen from 20% in Q1 2009 to less than 5% in Q1 2011. The reverse has happened for XML returns, rising from 12% in Q1 2009 to 27% in Q1 2011. CSO has established a specialist team, dedicated to assisting enterprises in making the transition from paper to XML in an effort to improve uptake.

5. Conclusion

No system is perfect; each has its own advantages and disadvantages. Some of the main pros and cons are outlined below:

Pros

1. The XML returns yield better quality data. Once the system has been set up properly, the questionnaire is interpreted properly every quarter providing consistent, clean data that require very few interventions;
2. The XML facility has contributed to a very significant reduction in burden for those enterprises availing of the system;
3. The XML project has provided positive publicity for CSO, portraying the office as willing to embrace new ideas and technologies.

Cons

1. The XML system is more difficult to manage. Coordination across multiple payroll software providers and mixed-mode data collection brings new challenges;
2. The XML system has introduced new (unforeseen) risks. For example, any changes to the EHECS questionnaire present a real cost to the software suppliers as they must programme the amendments. In the current economic climate, where enterprises are already under financial strain, there is no guarantee that all the payroll providers can make the required changes in the same timeframe. In addition when amendments are made to software (e.g. fixing a programming error) the software suppliers are dependent on their customers uploading the most up to date, corrected version of the software;
3. The XML has a high entry cost (or burden) for the enterprise, as the system requires exacting links to each person engaged;
4. The XML system introduces some loss of control. Everything must now be coordinated via the payroll software providers which can make change difficult. This also requires more frequent and careful communications.

Reflecting on what has and has not been accomplished to date with the payroll project there are a number of key lessons that have been learned (and no doubt there are many more still to be learned). These include:

1. *Always have a clear goal*. The focus of the initial project was not sufficiently clear. There were too many competing objectives with inadequate priority or ranking and as a consequence, the EHECS was allowed to become too ambitious and complicated. For example, the non-payroll data were desirable but un-necessary.
2. *Keep communications channels open*. Consultation and communications with all stakeholders is crucial to ensure buy-in. Respondents and IT systems providers will be more sympathetic if they understand what your objectives are and that ultimately, you are trying to help them.

3. *Keep the solution simple*. The majority of payroll companies provided the XML functionality for free by keeping it simple. Those companies who charged for the solution inevitably added unnecessary complexity. When changes or problems with bugs arose the simple systems were easily and quickly fixed. This was not always the case for the more complex solutions which led to additional cost to our respondents.
4. *Provide expert support from the start*. Prior to the review, support to enterprises on the XML functionality was provided on an ad-hoc basis. With the establishment of a dedicated, trained XML team to support and promote the functionality, the level of uptake is increasing each quarter.
5. *Promotion is critical*. Until recently CSO did not effectively promote the XML functionality with the result that many users were unaware of the facility. By promoting the functionality users can see that CSO is making efforts to reduce burden while taking into account their concerns and input. By sending promotional material in advance to all newly sampled enterprises together with follow-up calls from our XML support team we are encouraging these enterprises to set up the functionality in advance of a survey form issuing to them.

References:

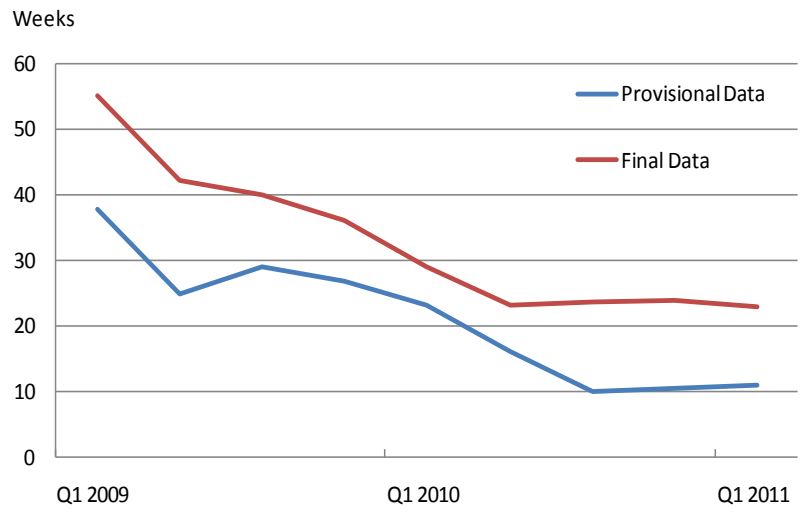
CSO (2010), "Review of Earning Statistics Collection and Dissemination in the CSO". Unpublished Management report, April 2010.

CSO (2011), "Response Burden Barometer - 2010". www.cso.ie

Appendix 1 – Management Metrics

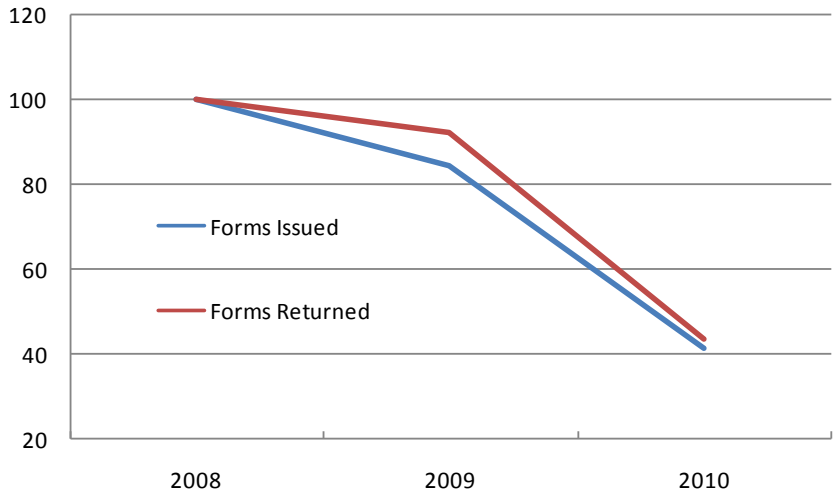
4.1 Timeliness

Figure 4.1 – Publication Time Lags Q1 2009 – Q1 2011 (Weeks)



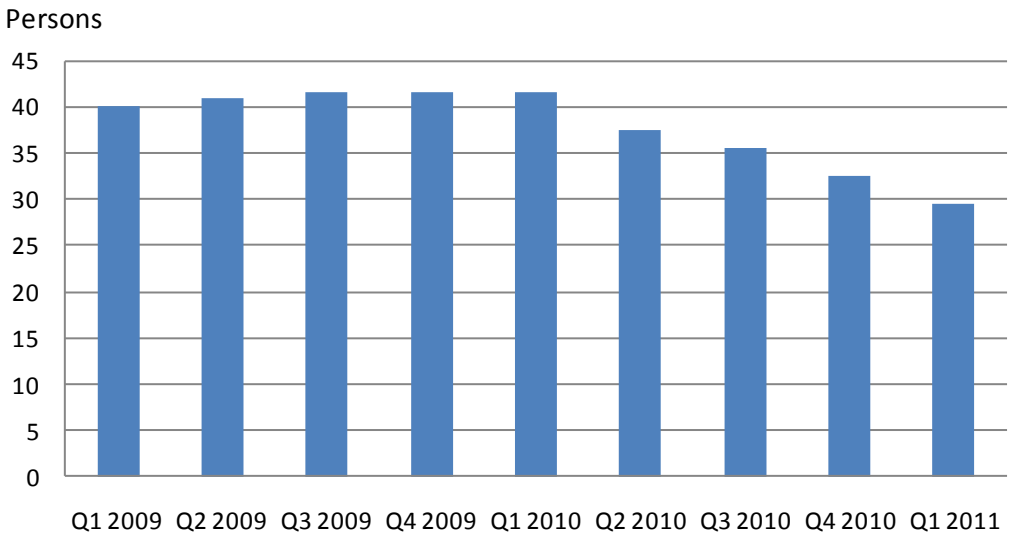
4.2 Burden Reduction

Figure 4.2 – Administrative Costs 2008 – 2010 (Index, Base: Year 2008 = 100)



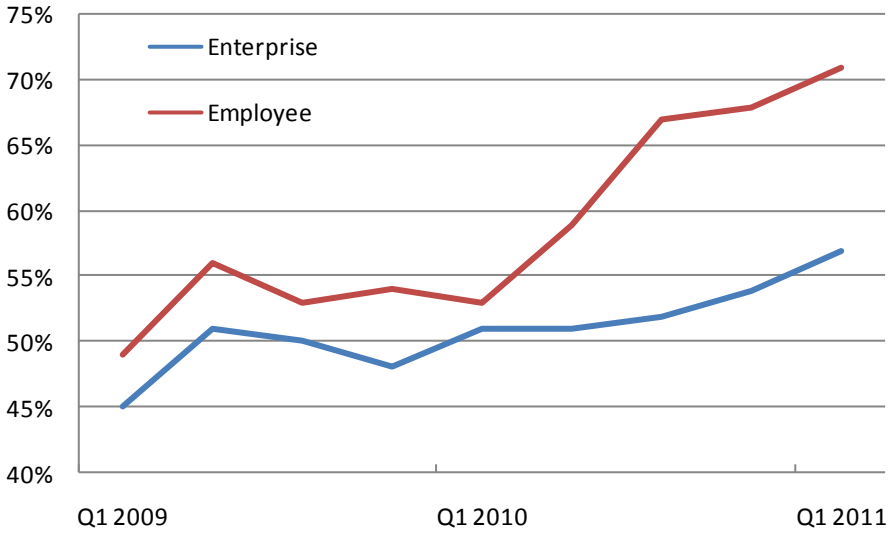
4.3 Resources

Figure 4.3 – Staff dedicated to EHECS, Q1 2009 – Q1 2011



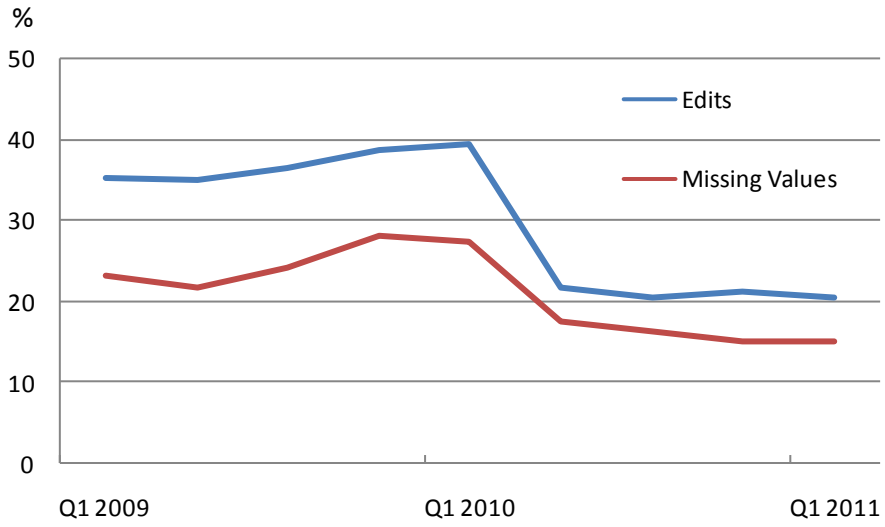
4.4 Response Rates

Figure 4.4 – Response Rates at T+70 days, Q1 2009 – Q1 2011



4.5 Required Edits

Figure 4.5 – Edits & Missing Values as a Percentage of Total Returns, Q1 2009 – Q1 2011



4.6 Share of XML returns

Figure 4.6 –Returns by Type, Q1 2009 – Q1 2011