Developing a digital tool for the Office for National Statistics to report on the UN Sustainable Development Goals

Ioannis Tsalamanis
Office for National Statistics, Newport, United Kingdom - ioannis.tsalamanis@ons.gov.uk

Nathan Eastwood
Office for National Statistics, Newport, United Kingdom - nathan.eastwood@ons.gov.uk

Abstract

The Office for National Statistics (ONS) has undertaken the task of reporting UK progress towards the UN Sustainable Development Goals (SDG) included in the "Transforming our world: the 2030 Agenda for Sustainable Development" on September 25th 2015. This meant that, by spring 2017, ONS needed a tool that would allow the collection, processing and dissemination of UK data for all national and global indicators specified by the UN. Such a tool should address short-term requirements for the ONS team but also take into consideration long-term functionality that would be necessary when the amount of data and usage is increased. For the short-term requirements, it was decided to invest in the development of an open source digital tool based on a Git repository hosting service that allowed the team to focus on the data identification, acquisition and storage. For the long-term requirements, it was agreed to develop a sophisticated web-based tool that would offer more stability, interactivity and visualisation capabilities. The team is currently using the short-term solution to prepare for the initial reporting on the SDGs and has been approached by similar institutions from other countries to share knowledge and expertise with them and help them develop similar tools.

Keywords: sustainable development goals, data collection, data dissemination, git repository.

1. Introduction

Between 2000 and 2015, the Millennium Development Goals (MDGs) motivated and mobilised the globe to achieve a set of important social priorities that would improve hunger, poverty, education, gender inequality, child mortality, disease and environmental sustainability and promote a global collaboration for development (UNITED NATIONS, 2015). The MDGs were geared more towards developing countries to improve their quality of life, and so the world's governments agreed to adopt a new round of goals that would continue this improvement and drive it to more sustainable practices for both developed and developing nations as well as business communities. After four years of negotiation, in September 2015, 193 UN member states ratified the Sustainable Development Goals (SDGs) to further tackle major world issues (UNITED NATIONS, 2016). With common goals covering economic, social and environmental targets, it is expected that they will broaden application and drive global change. The SDGs are expected to drive new behaviours and outcomes, and be achieved by 2030.

The SDGs include 17 goals which are underpinned by 169 targets that cover a broad range of current and future sustainable development issues. It was agreed that these targets are to be measured by 230 global indicators and, although some of these targets may not apply to the UK, it was agreed that the government should still strive to improve in some way. ONS is responsible for providing the data that can be used for monitoring and reporting on a global level, identifying the nationally relevant targets and where necessary, selecting relevant supplementary national indicators. ONS must track UK progress towards the SDGs by analysing the relevant data, identifying any data gaps within the indicators and suggest data sources or techniques that can address these data gaps. Finally, ONS has to develop a mechanism that would allow the efficient collection of all the relevant data and dissemination of the data using tables and appropriate

visualisation techniques.

The Data Science Campus (DSC) team, which sits within ONS, has agreed to collaborate with the SDG team for the development of the short-term solution and identified four different subsystems that need to be in place so that reporting on the SDG indicators can commence. The digital tool should offer metadata collection, data collection, data validation and dissemination. The metadata collection covers the collection of all relevant information about each indicator; this will include the data source(s), data availability and method of computation. The data collection covers the process of identifying the relevant data for each indicator and ingesting this data from both external and ONS sources. The data validation step will apply a series of business rules to ensure erroneous data is not published and the data provider is alerted in this event. Each of the reported indicators must be disaggregated by seven key groups: income, sex, race, ethnicity, migratory status, disability and geographic location; as well as any other relevant disaggregations. Finally, the dissemination process must be accessible to the general public online and incorporate several techniques of visualising the data, both tabular and graphical, whilst making them available in downloadable, common open formats. To achieve all the above, and to prioritise on the quick reporting on the indicators, the DSC team proposed the use of an open source, US built system which uses a GitHub's code repository hosting service in combination with their static, Jekyll based web-page hosting facilities called GitHub Pages, version control system and user authentication functionality.

2. Metadata collection and maintenance

For the first subsystem of the tool, the metadata should provide information for each of the 230 indicators and cover fields such as the indicator's definition; data availability; disaggregation information; data source(s); description of the process to acquire the data; description of the appropriate dissemination for each indicator; and other relevant information. The indicator's definition should include the official UN definition if the indicator is global and a custom detailed definition if the indicator is national. The data availability field should indicate the date the data for each indicator was acquired and identify possible data gaps and potential sources. The disaggregation field should contain information about the available levels of disaggregation for each indicator and identify possible limitations for each level. The data source field should distinguish the origin of the data and identify if the data is comes from external collaborators or belongs to the ONS. The acquisition process field should describe in detail the procedure that was followed to acquire the data from the source and the potential processing necessary to transform it to the required format for the dissemination system. The dissemination field should identify the tool used to visualise the information for each indicator and give a short description on how to use the dissemination tool to examine all the disaggregation levels.

By using the Jekyll web-page generator (JEKYLL, 2017) and GitHub (GITHUB, 2017) hosting capabilities of the Git repository service, the DSC team was able to quickly construct clone, modify and re-brand a US built website (U.S. General Services Administration, 2016) that could host all the UK SDG data and metadata. The data is stored in structured, flat file formats which will be easy to migrate should a more robust, long-term solution be required in the future, ensuring that time is not wasted. These fields are editable using a secondary service called prose (PROSE, 2016), which is linked to the GitHub account. prose offers a web-based interface that allows the user to create, edit or delete data without knowing or understanding the underlying programming languages. Ensuring a consistent approach to each metadata file allows easy integration with the dissemination system as metadata information can be easily accessed and used by the visualisation tools and other parts of the site.

Given the fact that the task of metadata collection and maintenance will be manual and given the high number of indicators to be processed, the high number of sources those indicators will likely map to and the variety of data providers, it is expected from the team that this part of the system would represent a significant amount of initial investigation work. In addition, data source(s) and other parts of the metadata are likely to change over time. To help with that, we use Git which is a version control system and allows controlled updating of the data, as well as the ability to view or rewind changes should the need arise. GitHub provides a way to host Git code repositories and extends its capabilities, for example with the use of authorised accounts, the administrators are able to give database access to specific users and control who can

add, remove or replace information and, more importantly, review any changes made by these contributors.

3. Data collection

Data collection covers the process of ingesting and disseminating data from both external and ONS sources. When the data comes from external providers the system should allow for authentication for trusted providers, ability to manually input data, upload a data file, verify the data and warn of problems with the format. Like the metadata information, for an external provider to upload data, they need to have an authenticated account with the GitHub repository and be in the providers' list of the administrator. Regarding the manual input of data, the tool should provide clear instructions to the providers as for the process that needs to be followed to successfully upload the data and details on the supported format and structure. For the data manager within the SDG team, it is important to have a view of the collection process, including peer review of input data, version control of data to track changes and data validation checks. For data whereby external providers are not able to provide the suitable level of disaggregation there will be a need for some data processing steps. In this situation key requirements for those performing the processing would be to have a fully documented workflow which would be separate to the tool.

All the requirements were met again by using the web-page generator and hosting capabilities powered by Jekyll and the authentication and version control offered within the GitHub service. Through working with some initial sample goals, the team was able to identify two different types of data sources; data that is available at the necessary level of disaggregation and data that needs some form of inference, aggregation or processing. Where the latter are not possible, this is referred to as a data gap. If the data gap falls in one of the existing business area remits within the ONS, the SDG team can speak with the relevant business area to explore how the data can be collected and processed. In this case, the collection and processing system to be used can be the decision of the business area. If it falls outside any business area remit then it is down to the SDG team to scope further resource, explore potential data sources and fund the collection and processing. In the long-term, since ONS has a dedicated Data as a Service (DaaS) team that is responsible for acquiring and processing data for the entire organisation, it is expected that they will help to undertake the task of collecting the necessary data as well as processing it to the appropriate format and structure for the dissemination tool. However, for the short-term, the SDG team is expected to collect the data themselves until the DaaS have the available resource to take on this role. As the data will only need to be updated once a year and come in a variety of file formats and structures ranging from flat file structures to PDFs it would be inefficient to attempt to design and implement a complex custom Extract, Transform, Load (ETL) system.

5. Dissemination

Data will need to be disseminated using tabular and graphical visualisations and be available for download in common open formats to raise public awareness. Data serving the website is in non-proprietary format that is ordered and stored in a structured way; but is also non-identifiable. The dissemination platform will need to provide capabilities of storing both statistics and metadata, restrict access using permissions and allow data providers to edit the data in a simple fashion. The DSC team developed an open source proof of concept system for both collecting and disseminating SDG data using a GitHub repository as an intermediate online platform.

Editing data capabilities are provided by the prose.io web service. Each edit of the data is known as a "commit" and creates what is known as a "pull request" which can be reviewed by the data manager who, when happy with the change, can "merge" that change to the repository. For each pull request submitted, owner users get notified via email about the proposed changes. These changes can then be discussed, reviewed and add follow-up commits before they are "merged" into the repository. Git allows you to revisit previous changes made to a repository and, by reviewing the log details, the user is able to switch the working repository to any point in history. Via the authentication and permission system of GitHub, the administrator of the repository can specify the level of access to the repository that the team can have. This ranges from full access to read only access. By building the entire project in GitHub, the team has the opportunity to access and manipulate it by various interfaces including a web front end, desktop programs and the command line. Repositories can be made private or publicly available depending on the sensitivity of the data used, but this

project will remain as open source and therefore publicly available.

GitHub Pages (GITHUB PAGES, 2017) are public web-pages hosted for free through GitHub and can be personal websites or websites related to specific GitHub projects. Pages is an extension of a GitHub repository, whereby if the Git "branch" is named in a certain way and files inside it are HTML or Markdown, the file(s) will be rendered as a static website. GitHub Pages is the self-aware version of GitHub and comes with the powerful static site generator Jekyll that was mentioned earlier. The repository is monitored for changes and when one occurs, the site is automatically rebuilt, ensuring it is always up to date with the latest data.

The approach followed for the short-term requirements of the dissemination system has a number of advantages. Initially, it is based on lightweight code that allows rapid development and users can be trained in a very short time. Data is stored as small flat files that are easy to manipulate and migrate to other systems. GitHub's authentication and peer review align well with the needs of the SDG team to develop a secure system that would have version control and registered accounts embedded. Moreover, excessive technical baggage is avoided and code and data are reusable and portable wherever possible. This solution sits well with the UN directives for open source, shareable solutions and could provide opportunity for collaboration on future development with other countries. One disadvantage of this solution is the bandwidth limitations placed on websites built using GitHub Pages which may force the SDG team to look for alternatives should the website have significant traffic.

6. Conclusions

To comply with the UN decision to work on Sustainable Development Goals, ONS had to come up with a quick solution on how to collect, process and disseminate information that could be used for reporting on the indicators related to these goals. Currently the ONS is in the process of developing a centralised data storage system that will give access to all the necessary data, but this service will not be available for usage until 2018. With this is mind, the development of the tool has been split into a short-term solution, where services that are currently available are used, and a long-term solution that will integrate with wider ONS services as they come online.

For the short-term solution, the SDG team collaborated with the Data Science Campus team for the development of a static website using open source digital tools. This prototype had to be developed in a short time-frame and be user friendly so that people with no high level knowledge and basic training could use it to upload and disseminate data. Moreover, it had to tackle the data volume that would be created by the number of indicators within the SDGs. There are 17 goals with 169 targets and 230 global indicators, disaggregated by up to seven groups; in addition to this there will be supplementary national indicators increasing the data volume further. This will require a significant amount of resource from the SDG team and therefore the tool should make this process as streamlined as possible.

Long-term, it is expected that more teams and business areas within ONS, together with external providers, will be delivering data via a more centralised and formal way in the form of a data lake. As data providers move to this rational delivering system, the SDG team will gradually hook its dissemination system to query this central pool of data. Using the experience gained from the short-term solution, it is anticipated that a custom website may be built to replace the current short-term solution; this custom site would be served by the ONS infrastructure and extract data directly from the data lake.

References

UNITED NATIONS (2015). United Nations Millennium Development Goals. Retrieved from: `http://www.un.org/millenniumgoals` (Accessed 15 February 2017).

UNITED NATIONS (2016). Sustainable Development Goals: 17 Goals To Transform Our World. Retrieved from: `http://www.un.org/sustainabledevelopment` (Accessed 15 February 2017).

JEKYLL (2017). Transform your plain text into static websites and blogs. Retrieved from: https://jekyllrb.com (Accessed 15 February 2017).

GITHUB (2017). How people build software. Retrieved from: https://github.com (Accessed 15 February 2017).

U.S. General Services Administration. Sustainable Development Goals Indicsators, (2016). GitHub repository: https://github.com/gsa/sdg-indicators (Accessed 15 February 2017).

PROSE, (2016). GitHub repository:https://github.com/prose/prose (Accessed 15 February 2017).

GITHUB PAGES (2017). Websites for you and your projects. Retrieved from: https://pages.github.com/ (Accessed 15 February 2017).