

Bringing OECD Data to Life

Terri Mitton, OECD

Introduction

In July 2014, the OECD unveiled a new website designed to make OECD data easier to find, use and understand. The new *OECD Data* portal (data.oecd.org) expands the reach of OECD data to non-expert and experts alike by providing a centralised gateway to all statistical content across the Organisation.

At the onset of the project, we knew that we needed to take innovative steps to make headline data available “digital first” to better satisfy the growing public demand for accessible and up-to-date data. We also needed to revolutionise access to OECD statistical information, making it possible for citizens and policy makers to be more informed when shaping policy. At the same time, the academic community is an important part of our audience so we had to make *OECD Data* compatible with scholarly publishing standards.

OECD Data represents a paradigm shift for the organisation, not only are we opening up our content to the world, we are encouraging people to actively redistribute it and embed it on their websites. To reach a larger audience of non-experts we are adapting the language used by specialists to terms used in mass media. We are also showing real-time data in new and engaging interactive visualisations so that complex data can be easily understood and more accessible. This paper looks at the key drivers and objectives of the project and the dimensions for implementation.

Objectives

In 2011, OECD conducted a Publishing Review which showed that significant improvements in online dissemination had been made in recent years. However, the Review highlighted that OECD’s publications and data were perceived as difficult to access and use. Consequently, the Organisation decided to accelerate efforts to make its publications and data more accessible and open.

Though the accessibility gap referred both to analytical and to statistical content, the main problem was found to lie in the way statistical data were presented. The study also noted the use of technical jargon made our content difficult for non-specialists to understand. The Secretariat proposed the solution, endorsed by Council, to develop a new, user-friendly data portal as a single, central gateway to access all OECD statistical data.

The main objectives of the **Data Portal** are to:

- **Improve findability of OECD data** by providing seamless and integrated discovery, search and navigation paths to published data across platforms
- **Improve the use and understanding of OECD data by a broader audience** by presenting key headline indicators in clear, accessible language
- **Increase the impact of data disseminated by OECD** by delivering a user-friendly data discovery portal presenting engaging interactive charts which can be easily shared

The portal has thus been designed to display high quality, accessible and consistent statistical content. It allows users to access OECD data and data-related content, and to reuse, embed and share this content, thus increasing the usage and impact of OECD statistical work.

The *OECD Factbook* and “*At a Glance*” publication titles already include a set of key indicators written for a non-expert audience. These indicators served as the basis of the initial set of data portal indicators. The text of the indicators needed to be reviewed to ensure that the language was understandable to a non-expert. To maximise discovery by a broad audience, the indicator titles are written in non-specialist terms which are most often used in Google searches, by media and blogs.

At the same time, the portal provides access to datasets which are available via an Application Programming Interface (API) so that applications can consume OECD data machine-to-machine.

Key dimensions for implementation

OECD is still struggling with a number of key challenges. These are strategic elements which need to be understood and addressed in order to maximise data dissemination. They include:

Technical challenges

Digital first real-time publishing

Historically, data was most often used as an input to static press releases and physical books instead of a ‘living’ output for use in interactive, online platforms.

Moving to “digital first” publishing model involved overhauling the architecture of OECD’s publishing systems to interlink a new MarkLogic bibliographic metadata and content repository with our existing web content management system, relational data warehouse, and online web platforms. What were once stovepipe systems are now interoperable and OECD indicators are semantically structured, laying the foundation for building future innovative outputs.

Real-time data are only usable when there are timely updates in the database. Data in publications are sometimes not up-to-date in the data warehouse at the time of publication release because data producers tend to focus on the press release and analytical publication which have more visibility than a database. It is necessary to streamline production processes to improve timeliness of updates.

Open data

Open is defined as machine-readable, ensuring that anyone (e.g. data distributors, media, NGOs, search engines, academia, etc.) can use this data in applications of their own design and build.

Openness is critical to ensure OECD data are found and used by third parties in a manner that is compatible with the expectations of the open data community.¹ In 2013, OECD undertook a study to understand the best practices of the open data community and to benchmark IGO open data practices. The outcome of this study was used to define OECD's open objectives and open criteria for datasets.

To achieve machine readability, OECD data must be stored in the central data warehouse, OECD.Stat, formatted ('curated') according to open-readiness criteria, and the OECD infrastructure must be scalable to match potential peak demands.

This requires significant effort across the Organisation:

- Data producers have to restructure existing published datasets into a machine readable format. With over 550 published datasets, this exercise is quite labour intensive.
- Some datasets need to be migrated into the data warehouse.
- New datasets must meet the open criteria for datasets.
- IT investments are necessary to scale for increased usage.

Organisational challenges

Use of plain language

Throughout the process of selecting and preparing data portal indicators, language accessibility has been a major focus.

The Publishing division has produced accessibility guidelines, which were reviewed by specialist 'plain language' consultants. During this process, it shared and compared ideas on accessibility and readability with representatives of other organisations or governments who have developed plain language guidelines. These guidelines are being applied to the data portal language to ensure the accessibility of the information presented. This requires that frequently used synonyms be substituted for some specialist terms.

The introduction of plain language has been a source of tension because the Organisation is much more comfortable using officially defined specialist terms.

Balancing quality with speed of publishing data

The main challenge is that data producers currently publish and update the statistical databases in the data warehouse, OECD.Stat, in a decentralised manner, without the benefit of a centralised open-readiness and quality check prior to publication. This creates a risk that the datasets made public have inconsistent quality and coherence, creating a poor user experience. To ensure full implementation of the Open policy, a workflow will be set-up to ensure the necessary quality and open-readiness checks before publication.

¹ See examples: <http://sunlightfoundation.com/opendataguidelines/>; <http://opengovdata.org/>

Publishing challenges

How do we reach audiences who do not come to our website?

From traditional publishing to social publishing

The project represents a profound shift in our publishing paradigm: we are letting go of some control and are accepting some risk. We are not only opening up our content to the world, we are encouraging people to actively redistribute it. However, we have built in more meaningful analytics that enable us a better understanding of how our content is being accessed and used. In addition, while embeddable charts can be re-published outside of our website, they remain dynamically dependant on our data warehouse so that they always present the latest available data from OECD. They are also branded with our logo and can redirect users to our website, thus reinforcing our brand as a trusted source of information. And all the while, OECD is more actively using social media to promote and enable audiences to engage and to participate.

Reaching the ecosystem of dissemination channels

OECD's daily online reach in the US is around 0.01% which puts us in the top 16,000 most-popular websites (Source: Alexa) Since, for most people, we're not "front-and-centre", we must actively promote our data and fight to increase awareness to take a larger share of users' time online.

Gardner and Inger² researched reader behaviour among academic and high-level student audiences. They found that academics start their research for scholarly information in eleven places. A post-it and hope strategy would miss 85% of the starting points and signposts used by this audience. OECD actively posts metadata and/or full text content into specialist research and other channels for published outputs. The result is 122,000 referrals to OECD a month, 24% of all traffic to our published content.

Other audiences are no different; they each have different online sites where they go to find information – this is especially true among professional audiences – law, finance/banking, tax – and more recently policymakers with the launch of Bloomberg Government last year aimed squarely at those within the Washington Beltway..

For non-professional audiences (citizens and civil society activists) we use consumer channels like Google Books, Amazon and Scribd.

And the results are good – around 360,000 publication readings a month via these non-OECD channels. Now publishing metadata for 'statistical data' is being integrated into these dissemination channels.

Sustainability of the model

Looking forward, the portal has to be able to evolve to improve search functionalities, take into account new types of content, the evolutions in users' expectations as well as the increasingly rapid technological changes. It also needs to continue to provide a unified and consistent user experience, an essential feature of accessibility. It is therefore critical that the portal is managed centrally, to ensure the maintenance of high standards of accessibility, quality and consistency for the full range of our target audiences. This would also serve to limit the duplication of effort in managing indicators across the Organisation today.

² From: Gardner and Inger (2012): How readers discover content in scholarly journals

The portal functionality could be extended with new features to improve the experience for external users and to enhance the efficiency of managing indicators. The portal content must continue to evolve to fill gaps and reflect different priorities.

The data portal presents opportunities to bring consistency and coherence from accessible core indicator content through to specialist indicators and related material across the Organisation. The data portal could be extended, allowing Directorates to manage their own indicators and promote them through their own web pages bearing in mind the wide range of needs across Directorates, while maintaining a common look and feel across the OECD website. The extended data portal would provide the tools and templates for Directorates to manage the wide range of indicator content currently managed on their own web pages in silos.

Directorates, with little visibility for their indicators on the website today could create a selection of indicators using the standard template to highlight their activities and share/embed charts. The data portal would link through to these pages. This would provide the ability to present indicators according to additional thematic groupings aligned with Directorate and specialist user requirements.

Indicators created by Directorates would become candidates for inclusion in the main indicators of the data portal.

Conclusion

Carrying out the plan to make data accessible and open has been challenging, but the main challenge remains ahead of us: we must ensure that the data portal is sustainable for the future to maximise dissemination of OECD's data.

The data portal and open API services were developed without any new financial contributions from OECD Members. Now that the portal has been developed, how will we sustain it for the future? This is still an open question. Finding contributions from Directorates is one way that is being investigated. Other options that could be considered are exploring funding or grants from foundations, sponsors or external sources.