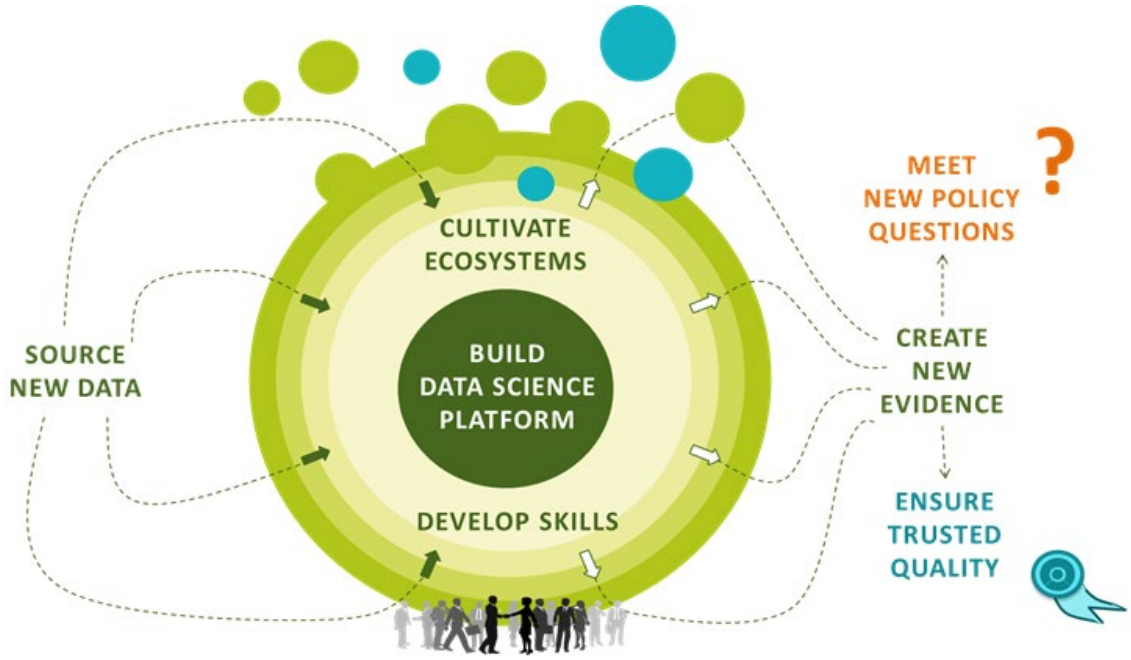
OECD Smart Data Strategy

Vision Statement

Developing sound policies in areas such as digitalisation, globalisation, sustainability, well-being, can only be done with solid evidence. Developing good evidence requires tapping into more data sources, continuing to modernise the existing data process, and leveraging advanced data science techniques, all while ensuring that the core value of providing trusted, quality evidence continues. More granular and timelier data help complement existing statistics, enabling micro-level analysis, improving forecasts and nowcasts. Also, combining sources enables multidimensional insight and modelling.

Developing new evidence to meet the policy demand

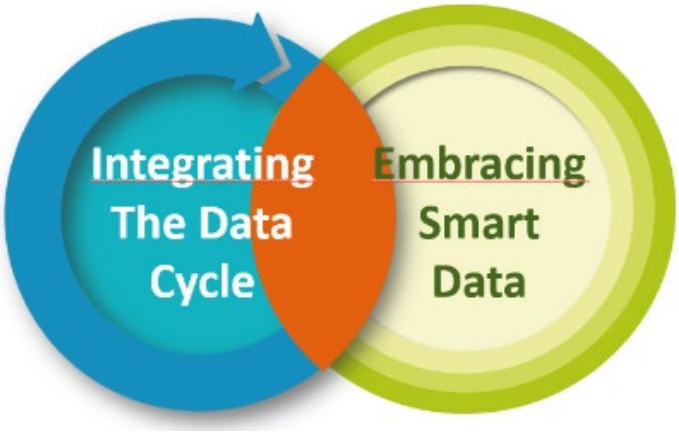


Meeting the policy demand hence requires to create new or augmented evidence, to complement existing one. This has implications in terms of capacity to tap into multiple new sources of data, from public and, more and more, private sectors; develop the skills and tools to combine the data, develop the models and analysis, whilst continuing to ensure trusted quality of the evidence upon which policy decisions can be made. At the OECD, we have identified more than a hundred of such innovative projects, involving new data sources and/or applying data science techniques in an innovative way in the policy field. One distinctive feature of such projects is that they can never be executed alone – by one team or even by one organisation – due to resource limitation, impossibility to access the relevant data, or necessity to pool in various expertise.

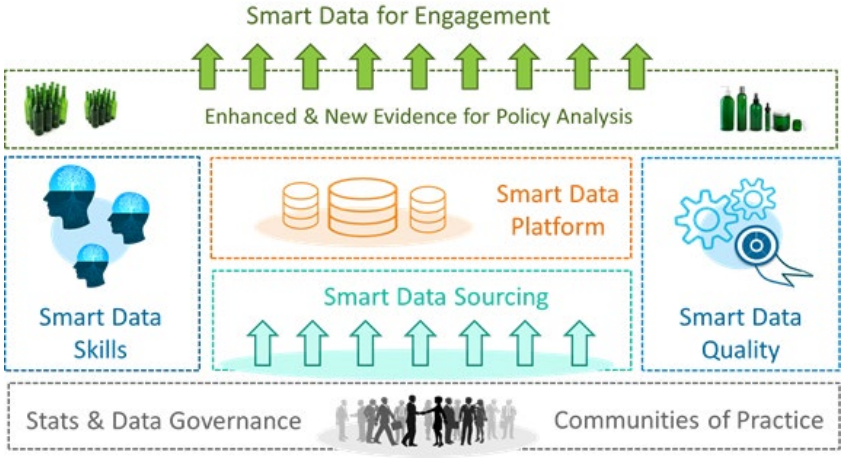
Smart Data requires us to think and act as a part of the broader data ecosystem, and hence actively cultivate and invest in this ecosystem.

Indeed, meeting the policy demand with data innovation requires continuous investment in data commons, including developing technical, organisational, legal, and human capabilities. The ***OECD Smart Data strategy*** pursues the development of such capabilities, in close collaboration with OECD members countries and the broader data ecosystem, with two broad goals in mind:

- **Integrating the Data Cycle**: modernising the collection, analysis, processing, dissemination of data, mainly fed by established sources – countries' statistical or policy reporting. This line of action is focused on increasing the efficiency in data operations, by overcoming fragmentation in tools, processes and data models, and enforcing a "quality by design" approach; harmonisation of data models is the cornerstone enabling data integration, efficiency and accessibility.
- **Embracing Smart Data:** tapping into alternative data sources and mainstreaming data science techniques such as machine learning or text mining. This line of action is focused on augmenting existing evidence (nowcasting) by bringing more granularity (spatially, or according to criteria such as gender or income distribution) and higher frequency, as well as bringing about new correlations and predictors that enable deeper analysis or improved simulation of policy effects and forecasts.

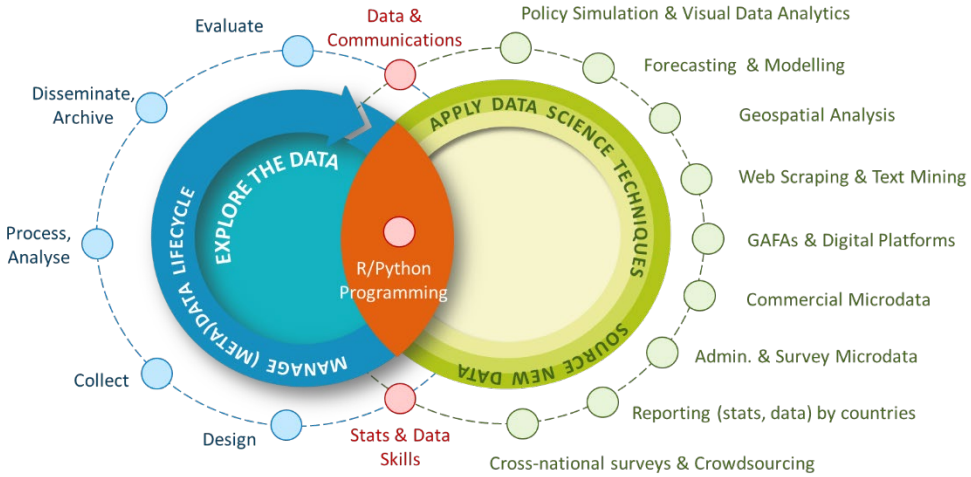


To achieve these goals, the *OECD Smart Data Strategy* identifies 6 interconnected data commons where innovation and investment are to take place:



1. The OECD Statistics & Data governance, under the leadership of the OECD Chief Statistician, has been reshaped, to appropriately mobilise and guide the community of data producers and data analysts, working in decentralised teams, close to each policy area. The new governance aims for increased coordination, joint investment in capabilities and cross-fertilisation, including through **communities of practice**, bringing together experts from all policy domains to solve common issues and mainstream emerging innovative practices.

Communities of practice to mobilise expertise & mainstream innovation

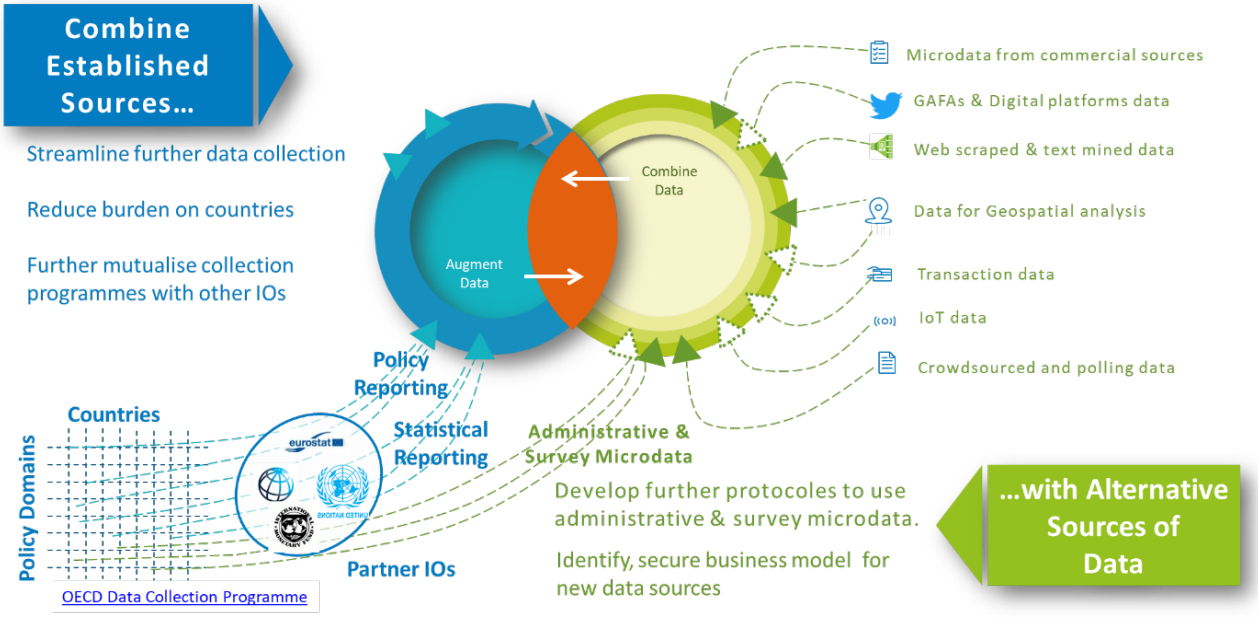


There are currently a dozen active communities of practice, addressing topics such as: geospatial analysis, algorithmic design, administrative micro-data analysis, policy simulation, data modelling, forecasting techniques, etc. More and more these communities are connected to similar communities in other organisations. Connecting communities of practice could be a powerful lever for knowledge spill over, data sharing, cost mutualisation, joint R & D and innovation, or indeed possibly increased collective weight in the dialogue with private data providers (see next section on data sourcing).

2. The Smart Data sourcing strategy identifies challenges and lines of action in three directions to combine established sources (statistical and policy reporting by countries) with alternative sources of data:

a modernising the engagement with *public data providers* (enhanced, more granular and efficient reporting; agreement to develop new indicators; secure access to or usage of administrative and survey micro-data for the purpose of policy analysis);
b strategic engagement with *private data providers* (according either to commercial or non-profit data sharing models) – notably, in partnership with digital platforms ('GAFAs') and suppliers of 'big data' in general, from every sector of the economy; and
c expanding *direct data sourcing* through webscraping and crowdsourcing techniques – both potentially mobilising social players and NGOs, and ultimately reaching out to citizens and/or specific business communities so that they can contribute their data to the public good.

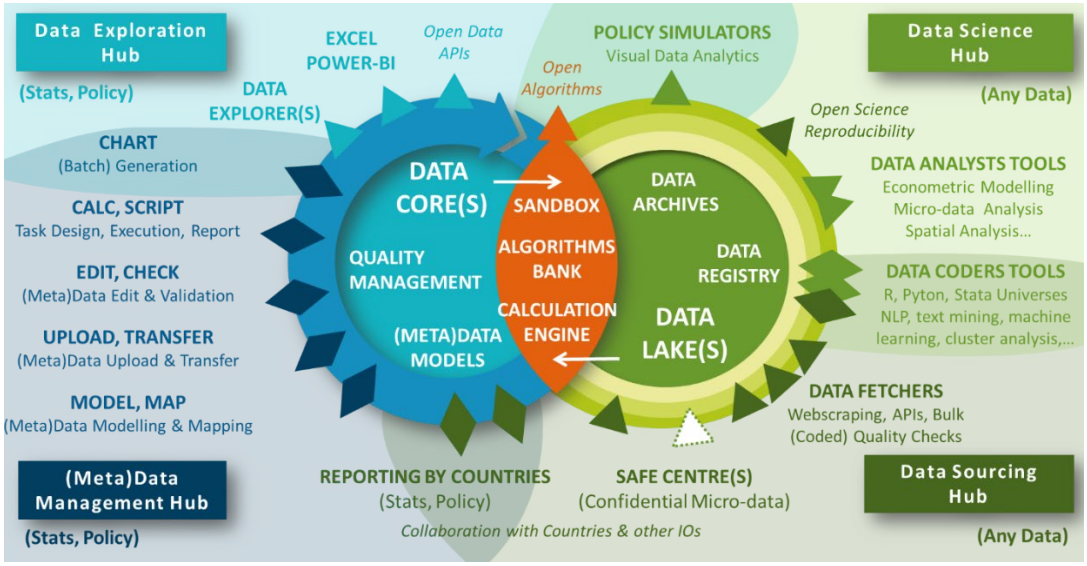
Smart Data sourcing: Combine establish sources with alternative ones



3. The Smart Data platform comprises 4 functional hubs, to rely on a powerful distributed computing and storage IT infrastructure:

a the *data sourcing hub*, for data producers to store large amounts of structured or unstructured data ("data lake"), to share data sources with their peers ("data registry"), to automate and de-duplicate data fetching and cleaning tasks, and to securely exploit confidential micro-data sources ("safe centre");
b the *(meta)data management hub*, for data producers to manage data output production over the entire data cycle, integrated thanks to harmonised data models (SDMX) and shared algorithms ("algorithms bank");
c the *data science hub*, for data analysts to quickly explore, combine multi-source data and develop early insights, elaborate quickly models and simulations; and
d the *data exploration hub*, to support basic dissemination services – exploration of the database to retrieve data of interest in the relevant format, connection to APIs for machine-to-machine data consumption.

Smart Data platform: Towards a more integrated and open platform



The *.Stat Suite* open source, SDMX-native solution, developed with the SIS-CC community is the cornerstone of the *OECD Smart Data platform*. It is now supported and financed by 15 national and international organisations, and is being deployed progressively in lower-middle income countries as part programmes to upgrade their statistical infrastructure – see SIS-CC 2020-25 strategy.

4. The Smart Data quality framework, capitalises on the well-established OECD data quality framework, to expand in several directions:

a *renewed quality objectives* (intrinsic data quality, timeliness, accessibility, reproducibility, security-privacy), to be captured ultimately in a data quality contract agreed with stakeholders/providers, and consumers of the data product – and enriching the existing framework with contributions from the field of open science and IT security;
b extend the framework to *alternative data sources* and address quality issues specific to each type of data – the challenges here being to bring about consistent and comparable quality across heterogeneous and diverse data sources not originally designed for policy analysis;
c apply quality approach to *all stages of the data cycle* (design, collect, process, analyse, disseminate, evaluate); and
d complement the institutional quality review process with more bottom-up, *ongoing quality improvement* approaches (driven by communities of practice, or through more systematic peer-reviews, for example).

5. The Smart Data skills framework identifies the needs for upgrading the skills in the Statistics & Data community (and 3 sub-populations: *data analysts*, *data producers* and *data engineers* – each featuring part of the competences of the 'data scientist', including *data coding* skills or the mastering of AI techniques), at different career milestones: recruitment (job profile), on-boarding of new staff, continuously training (leveraging online self-learning, data coaching, certification path and communities of practice) to upgrade data skills and raise the level awareness on data innovations.

The skills framework should also better assess the strategy in terms of in/out-sourcing, as well as the options for a multi-facetted partnership with a few key academic partners specialised in data science education.

6. Smart Data for engagement identifies the new landscape in terms of accessible data for policy products, and new ways of communicating and engaging on data with various audiences. In the post-truth era, making trusted data and facts accessible and bringing them to the public debate is a fundamental added value for producers of public statistics and data. To address this challenge, the strategy consists in:

a establishing a consistent multi-chanel engagement by *renewing the OECD web-ecosystem dedicated to data dissemination*;
b creation of *an extended layer of data indicators on horizontal themes*, and shifting from flagship publication paradigm to digital reports with interactive data; and
c *impact measurement*, to monitor and evaluate the levels of engagement, usage and impact generated across OECD and non-OECD channels.

Open ecosystems for data innovation

Smart Data is also about cultivating and connecting open ecosystems for data innovation, connecting experts from national and international public organisations, as well as academia, start-ups, NGOs, large digital players, sharing and co-constructing intangible assets (open source, open data, open algorithms and open knowledge) as public goods, in order to design better, evidence-based, policies. **Pivoting to an ecosystem-driven model requires a shift in the 'business model', as well as an upgrade of the data governance framework**:

➔ New 'business models' for data production and innovation need to be invented, that entail a much larger degree of co-investment in data capabilities, specifically geared to lower-middle income countries , and a much broader sharing of data and knowledge. Successful initiatives exist, that we could draw on to achieve much larger ambitions – such as the WB-led Development Data Partnership of which OECD is a founding member, UN Global Platform, or the OECD-led Statistical Information System Collaboration Community (SIS-CC)[1].

➔ The level playing field for data needs to be redefined, including in order to address multiple challenges and frictions that are orthogonal to the open ecosystem paradigm: in particular with regards to *privacy protection*, *competition between market players*, or *cybersecurity threats*. In the broad data governance agenda, the 'smart data for policy' angle should be nurtured so that public and private providers adopt the mindset, practices and technologies that should enable an efficient use of their data to better measure socio-economic phenomena, and ultimately design better policies for better lives.

For more information on the OECD Smart Data strategy, contact: OECD Statistics & Data governance secretariat; more on smart data strategies in this report: Which Strategies for NSOs in the Digital Era? Towards 'Smart Data' Strategies.

[1] "*SIS-CC is a reference open source community for official statistics, focusing on product excellence and delivering concrete solutions to common problems through co-investment and co-innovation*" (source: SIS-CC 2020-25 Strategy).