OPEN FOR FEEDBACK

Data institutions in China: stewarding data in the cultural and migrant population sectors

ODI Fellowships: report

February 2023

Open Data Institute

Contents

About

This report was produced as part of the ODI research fellow scheme. Its author was Dr Feng Gao, Executive Director of Open Data China, with contributions from Joe Massey, Dr Jared Keller and Jack Hardinges, ODI.

If you want to share feedback by email or would like to get in touch, contact Feng at gaoofeng@opendatachina.org.

If you would like to learn more about the ODI research fellow scheme, please visit the information and application page, or contact us at fellowships@theodi.org.

ODI Fellowships

This report is authored by an ODI fellow. It draws on concepts developed by the ODI but the author's views are their own.

Executive summary

This report explores data institutions in China through two case studies: one of an organisation stewarding national survey data and making it downloadable to approved users, and one of a public library stewarding linked data extracted from the collections of books and multimedia made accessible through an open API.

Introduction

In 2020, the China State Council announced an important policy,[1] listing data as a key factor of production – joining other factors which include land, labour, capital, and technology. To establish and fully grow the data market, the Chinese government has intervened in different ways to accelerate the flow of data, help the formation of data intermediary markets, and maximise the impact of data applications.

Open government data is generally considered an effective method to unlock the value of data by developed and developing countries, international organisations (OECD and World Bank) and renowned consulting firms (Mckinsey). It is now widely adopted and implemented across the globe, resources like the Global Data Barometer and the previous Open Data Barometer can be referred to for more information on global practice and performance.

Since 2009, China has introduced and explored open government data using a bottom-up approach where cities and provinces took initiatives before the national government to establish local open data portals and create local policy frameworks. Shanghai, for instance, developed one of the earliest local initiatives. It conducted feasibility research in late 2009, launched the local portal in 2011, and released a dedicated local policy framework in 2016[2]. After over 10 years of development, local open government data platforms have grown from several instances to hundreds[3][4], covering all levels of government in China with a major presence in developed regions. But behind this growth is a tension with the logic of open government data in China: open government data originally lays its legal foundation on the logic of transparency and accountability, while in China the logic of economic growth actually guides and drives the implementation of open government data.

When open government data was first introduced into China in 2008 (also known as 'Open Government Information' (信息公开)), it was considered a natural extension of freedom of information[5]. Open government data, therefore, was framed as a duty of a government agency. In practice, however, such duty is not accompanied by any mechanism to recover any costs associated with data cleaning and processing. Furthermore, there is also no incentive for government agencies to meet the data needs of the data market. Thus Chinese governments face bottlenecks in opening up high-quality data: the current rules and laws built upon transparency logic can not help move forward an economic-oriented open data agenda as government agencies are not willing and have no capacities and resources to satisfy the market needs[6].

---

[1] Communist Party of China (2020), 'Opinions of the Central Committee of the Communist Party of China and the State Council on Constructing a More Complete System and Mechanism for the Market-oriented Allocation of Factors'

[2] Shanghai Municipal People's Government (2018), 'Administrative Measures of Shanghai Municipality on Public Data and All-in-One Netcom'

[3] Fudan University Digital and Mobile Governance Lab (2020) 'China Open Data Index'

[4] Global Data Barometer (2021), 'Global Data Baromete: China'

[5] People's Republic of China (2007), 'Regulations of the People's Republic of China on the Disclosure of Government Information' (English Translation)

[6] Dr Feng Gao (2021), 'ODI Fridays: Open data and China – a ten year review'

Given the above challenges, local governments in China have made efforts to redesign their rules and systems to enable high-quality data flow. Some pioneering local governments such as Shanghai and Zhejiang Province attempted to redefine open data[7] as the action of making internal data accessible and usable to the public.

They expanded the concept of open data to cover 'non-conditional open data' which refers to the original 'open data' itself and 'conditional open data' which refers to the more general concept of data sharing, including public access, group-based access and named access as described in the data spectrum by the Open Data Institute (ODI). 'Conditional open data' can employ measures to control access, such as requiring an application to access the data; accessing data through a controlled environment; or requiring a fee to use data. It seems that such a redefinition of open data has already been widely accepted across the country, as Shandong Province[8] and Shenzhen city[9], for example, soon adopted the same definition in their local policy frameworks for open government data. At the same time, to address the problem that government agencies may lack the capacity and resources to deliver high-quality data, some local governments such as Beijing and Shanghai are further exploring how to delegate one or many authorised institutions[10] to take the responsibility for data stewardship. These organisations are charged with making data accessible – either fully open or conditional open – to the public or selected users to help realise the data value.

I found the above idea resonates with the ODI's concept of data institutions which are organisations that steward data on behalf of others, often towards public, educational or charitable aims. The practice of delegating stewardship of data to third-party institutions is not completely new in China but is often not well-known nor studied through the lens of data institutions. These practices were called or referred to in the news as data labs, data collaborations, or data platforms.

As 'data institution' is an umbrella term that includes various alternative data governance models, I use data institutions as a concept to investigate different practices in China to unpack their similarities and differences. I have curated a China Data Institutions register covering those example practices based upon the scheme of the ODI's Data Institution register. Those institutions can be studied in detail to inform local governments in China how to further develop rules and laws to enable the delegation of data stewardship. Research around Chinese data institutions will also enrich the existing ODI data institution case studies by adding examples from China and by comparing international practices.

I also conducted preliminary case studies by selecting two data institution examples from the register (as presented below in the table). These two institutions are both publicly funded but one is part of a government agency and one is a public cultural institution. Both of them have more publicly available information than other institutions in the register. I worked with these two

[7] Shanghai Municipal People's Government (2018), 'Administrative Measures of Shanghai Municipality on Public Data and All-in-One Netcom' ("上海市公共数据和一网通办管理办法")
[8] Shandong Province People's Government (2022), 'Administrative Measures of Shandong Province on Open Public Data' ('山东省公共数据开放办法')
[9] Shenzhen Municipal People's Government Regulations (2021), 'Shenzhen Special Economic Zone Data Regulations' ("深圳市数据条例")
[10] Such practice is coined as "授权运营". The earliest example is found in Chengdu (in Chinese) where the city government authorised the Chengdu Big Data Company which is state-owned to steward all government data assets and make them ready for external users to access and use.

institutions during a competition programme, SODA (Shanghai Open Data Apps),[11] and these institutions served as data partners to make data accessible to competition participants. Therefore I could use meeting notes and direct observations about how those institutions actually worked to complement the public information about them.

I investigated the two case studies mainly through desk research by researching materials such as websites, publicly available talks and presentations, media reports as well as past meeting notes. I also conducted expert interviews to complement the desk research. The Data Ecosystem Mapping tool as well as the data-use journey framework and the four levers for facilitating safe access to sensitive data framework are employed in analysing and illustrating the details of each case study.

Data Institution	Description
Migrant Population Service Center	Migrant Population Service Center is an agency under the supervision of the China National Health Commission. Since 2014, it has stewarded the China Migrants Dynamic Survey Data (CMDS) and made it accessible to external organisations through the Migrant Population Data Platform. The CMDS contains migrant population information collected annually from 2009 to 2018, including but not limited to basic population information, mobility trends, public service supply, and other multi-dimensional survey data. Organisations can apply to access and use data either individually or grouped by year, region, province, or theme.
Shanghai Library	Shanghai Library stewards linked data and identifiers about names, locations, historic events, and cultural works that are extracted and produced from Shanghai Library's rich collection of books, journals, photos, and multimedia. The initiative started operation in 2016 and makes all data resources accessible to the public through open API and SPARQL endpoint under a Creative Commons licence. The data institution has been running an annual competition since 2016 to make data useful not only for professionals but also for ordinary citizens.

[11] You can find more information about SODA on how it was designed and operated in these two-part articles on Paris Innovation Review: Part 1 and Part 2

Migrant Population Service Center

Migrant Population Service Center is an agency under the supervision of the China National Health Commission. It stewards the China Migrants Dynamic Survey Data (CMDS) and since 2014, has made it accessible to external organisations through the Migrant Population Data Platform. The CMDS contains migrant population information collected annually from 2009 to 2018, including but not limited to basic population information, mobility trends, public service supply, and other multi-dimensional survey data. Organisations can apply to access and use data either individually or grouped by year, region, province, or theme.

What data does MPSC steward?

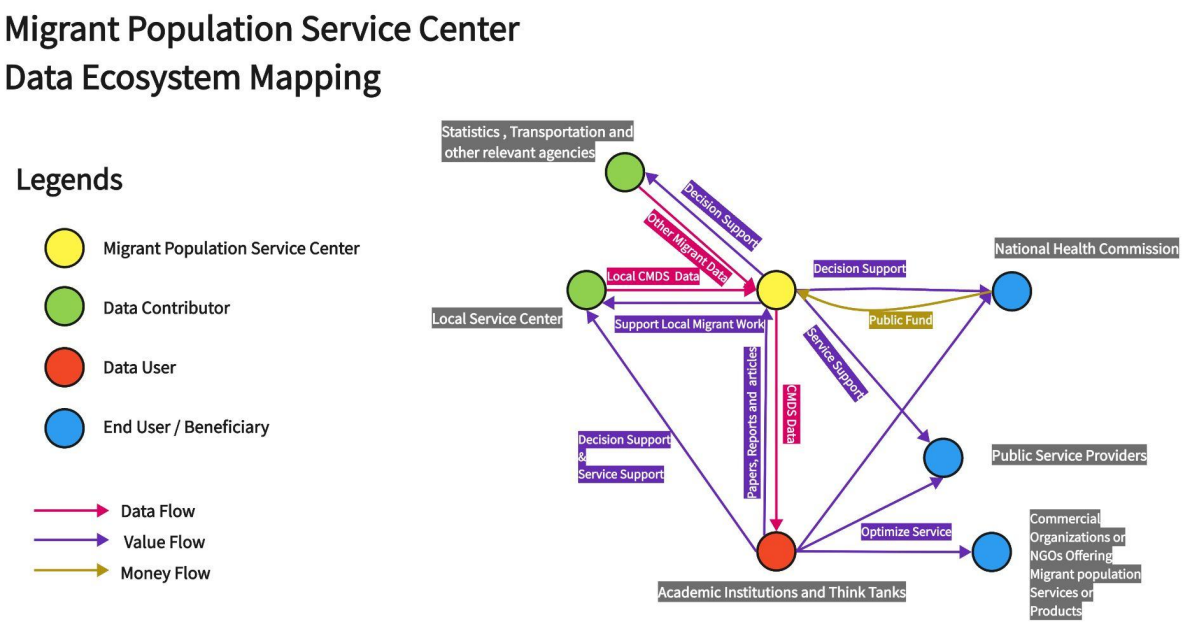

Figure 1: *Migrant Population Service Center Data Ecosystem Map. Ecosystem map showing the relationship between the Migrant Population Service Center (in the centre of the diagram) and: statistics, transportation and other relevant agencies; local service centres; the National Health Commission; academic institutions and think tanks; commercial organisations or NGOs offering migrant population services or products. The dataflows include other migrant data, and local CMDS data. The value flows include decision support; service support; papers, reports and articles; and support local migrant work.The money flow shows 'public fund'.*

Since 2009, the National Health and Health Commission has conducted the CMDS – an annual large-scale national migrants population survey. The survey is focused on the so-called migrant population or floating population, who are

without local household registration status through the Chinese Hukou system[12] but currently work or live in the country. The survey covers 31 provinces (autonomous regions and municipalities) across the country and the areas where the migrant population is relatively concentrated. The annual sample size is nearly 200,000 households. A standard questionnaire is used to collect information from each household including basic household information, income and expenditure, psychological status as well as their access to public services such as health services, marriage and family planning services, and education. In addition, it also includes a special survey on social integration and mental health of the migrant population; a special survey on health and family planning services in outflow areas; and a special survey on medical and health services for the floating elderly.

The individual responses to the survey are collected and processed into structured data: China Migrants Dynamic Survey Data. The data can be further grouped by years, regions, provinces, or themes such as mental health status. The Migrant Population Service Center conducted its own analysis and reported the results back to the National Health Commission.

Why make data accessible?

The most important driving force to make the CMDS accessible comes from the desire to maximise the data's value. Considering that the CMDS contains complex and rich information about individual households, the National Health Commission recognised the difficulty of fully realising the value of the data on its own. Therefore, in 2014, it decided to pilot the open sharing of the CMDS to bring in 'external brains' to maximise the value of CMDS.

At the same time, the Migrant Population Service Center, which had just been established, was also seeking its own differentiated positioning as one of the institutions directly under the National Health Commission. Driven by the opportunity, it was appointed as the executive unit to make the CMDS accessible to external organisations. The centre accepts applications from research institutions and think tanks for data access every year during a specific time window. The process is already standardised into five steps:

1. an individual registers an account
2. the institution the individual represents/ belongs to files and certifies the application
3. all applications are reviewed internally for eligibility checks
4. data agreements are signed and data is provided for download
5. the individual should report any outputs of data use via the online system.

During its seven years of operation (from 2014 to 2021), the Migrant Population Data Platform has had more than 600,000 visits[13] and processed applications from more than 7,000 individual experts representing more than 230 organisations. The use of the CMDS produced more than 1,500 academic articles and helped multi-level government entities reshape their policy on the migrant population. The preliminary estimate of the social and economic value exceeds 200 million yuan ($31m, £22m).

[12] Hukou system is a household registration system used in mainland China.
[13] Figures are calculated or estimated based upon a public talk delivered by Prof Ying at Fudan University (2021) who led the development of the data platform for the Migrant Population Service Center.

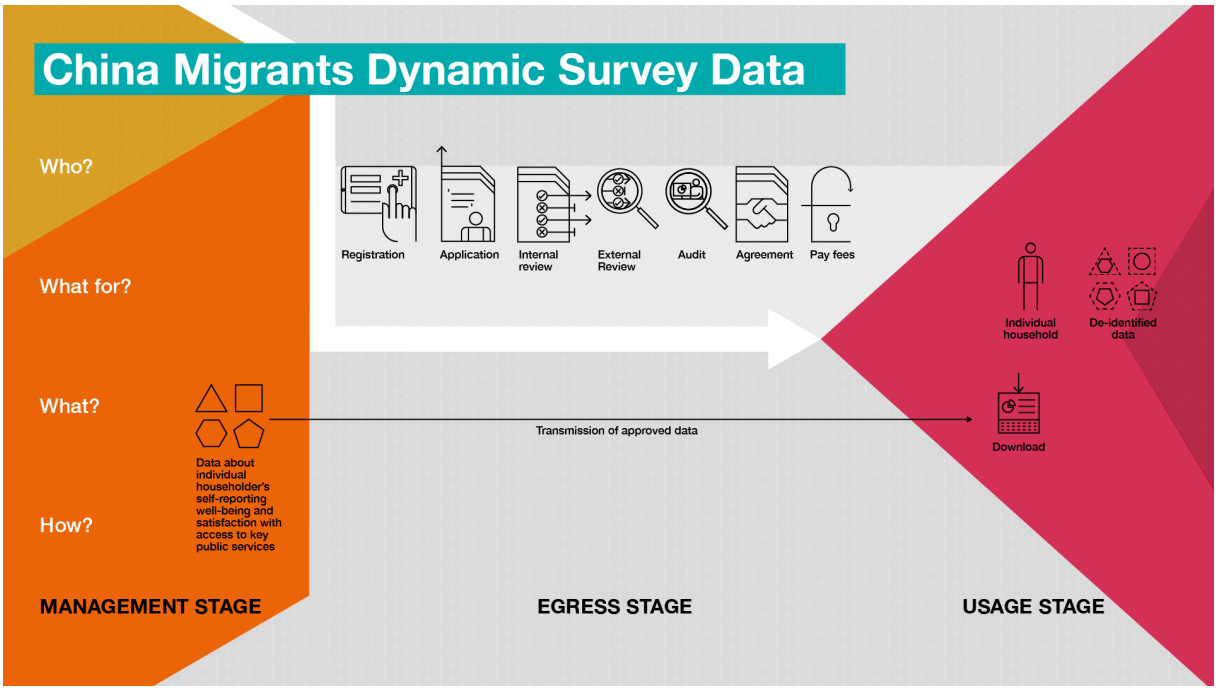

Figure 2: *How the Migrant Population Service Center facilitates safe access to data. China Migrants Dynamic Survey Data. On the left in the management stage, asking: who?, what for?, what? and how?. In the centre egress stage, the diagram includes registration; application; internal review; external review; audit; agreement; and pay fees. The egress stage includes 'transmission of approved data'. The usage stage on the right includes individual household, de-identified data and download.*

How to facilitate safe access to data?

The Migrant Population Service Center has a standard process and employs various safe access measures to ensure the data can be safely shared with the right people for the right purpose. The process and associated measures are explained below:

Who can access the data?

The Migrant Population Service Center makes it clear that any individual can register an account on the Migrant Population Data Platform, but to apply for data access, the individual must be affiliated with an organisation of any type and size, so the application form and data agreement can be stamped by the organisation. This requirement helps the centre ensure that the data users are traceable and can be held accountable through affiliated organisations, while the centre itself does not need to invest in heavy resources to directly manage a large number of individual data users.

Application process

Annual application window

The centre currently only accepts applications within a specified annual application window, which usually is about one month. The purpose of this window is mainly to reduce the uncertainty of staff resourcing and to focus on efficiently collecting and processing data applications within a limited time window.

Structured application form
Applicants are required to clarify the purpose of use and state the expected output of the data usage in their application form. And it is important for applicants to specify what data they are applying for: one can choose which year, which one or multiple regions/provinces, and which topic (e.g. mental health, access to education etc) the data is about. Last but most important is the application form must be approved and stamped by the organisation the applicant is affiliated with.

New applicant v.s. returning applicant
It should be noted that, in order to encourage data applicants to report back any outputs of data use in a timely manner, the centre only allows first-time applicants to request data of a single year, or data covering a single region/province, or data about a single topic while returning applicants who reported outputs are rewarded to request the whole collection of CMDS. In case an applicant fails to report any outputs, then the applicant will not be approved for future applications.

Application review

Format rather than content
The purpose of an application review is to check the completeness of the application form and validate it is stamped by an organisation so that later a legal entity can be held accountable. It is not, however, a process to make any judgement on the proposal of data use and only grant data access to selected best proposals. In other words, it does not matter what you propose to do: be it a research study or a visualisation analysis,it is irrelevant to the result of the review. As such application review can be simply done by the staff team at the centre, it does not require any experts' input. However, it is worth noting that this simplified review process takes the risk of granting data access to some unethical proposals, as the centre currently assumes that the applicant’s affiliated organisation is responsible for conducting the ethical review internally. It would be better if the centre can introduce a more rigorous process (eg conducting its own ethical review or requiring proof of passing an external organisation's ethical review) to ensure responsible use of CMDS.

Approve as many applications as possible
It may be surprising to hear that the principle guiding the review process is "approve as many applications as possible". Whenever the review team finds problems in an application, the team will actually proactively reach out and guide the applicant on how to improve the application. It helps both parties save energy and time in getting the application paperwork done correctly.

Final approval by National Health Commission
After the centre has screened all applications and made a long list for approval, it will submit the list to its supervisor, the Migrant Population Division at the National Health Commission, for final approval. The process seems a bit like how a visa application works: an applicant needs to submit an application to the visa centre. The visa centre is responsible for screening and reviewing applications and then submitting them to the Embassy for final approval.

Data anonymisation and transmission

After being approved by the centre, the applicant can now login to the platform and download the requested data file. The data file contains each individual household’s responses to the CMDS survey questions. This covers either the full

set of questions or only the selected questions related to the topic as requested by the applicant, such as mental health or education.

It should be noted that the original CMDS contains information that could potentially be used to identify the specific household including the household's address, and each member's basic information eg birth date, gender and marriage status. The data file for download is a de-identified version of CMDS which removes the address and converts individual member's basic information into aggregated statistics describing the household.

Training and peer learning network

After the review process, how can the centre ensure that data users can truly understand the data and make good use of the data? The centre creates an interesting strategic event to both train data users but also create a space for data users and data provider teams to know better about each other and build connections.

The so-called "Forum on Sustainable and Health Development for Migrant Population" invites all successful applicants to attend the forum but also invites experts and relevant officials from the National Health Commission to give talks.

Experts and officials from the Commission usually deliver training-oriented talks to educate successful applicants about the background of the CMDS and share what are the most pressing issues regarding the Migrant Population the Commission expect to address. Successful applicants are also given opportunities to give talks on what they are going to do as well as showcase what they have already done regarding the Migrant Population issues. Therefore, an expert peer learning network on Migrant Population issues is created through the open sharing of CMDS.

This forum also helps the Commission itself to identify and be connected with new experts, especially young scholars, and front-line practitioners. They will be engaged in not only realising the value of CMDS but also contributing to the Migrant Population policy agenda in other possible ways.

Highlights and reflections

The centre faces two major challenges in operating the CMDS:

Cultivate a layered user community

The main user group of CMDS is a group of researchers investigating the relevant issues facing migrant populations. The Migrant Population Service Center attracts those researchers through a competition process (data application process) but engages them not only through offering data access but also through offering networking opportunities as well as connecting them with the relevant level of government agencies to support decision-making.

In addition to this research group, the centre also wishes to attract non-academic organisations such as NGOs and commercial companies that can use CMDS data to optimise their charity programmes or professional services to help migrant populations. But the challenge for the centre is how it can follow up and measure

the impacts those non-academic organisations may make. The potential impact could be less quantifiable compared to the number of papers.

The centre will need to develop new capacities and build new partnerships to follow up on such practical usage of CMDS and evaluate the impacts. The centre also should partner with third-party data platforms or data search engines to increase the visibility of CMDS so that wider user groups can be attracted, identified, and engaged.

Towards data collaboration: how data access stimulates data contribution and exchange

The centre understands the CMDS is limited to sample-based questionnaires while there are other types of data potentially coming from other statistics, surveys, or even real-time data sources such as apps or sensors that can help enhance the understanding and research of the migrant population. For instance, it could be interesting to access aggregated or even individual anonymised shopping data (for example, goods, types of goods, price, and frequency) from large e-commerce sites such as JD.com or Taobao of Alibaba. Such data will be able to provide additional evidence or indicators to assess the wellbeing of the migrant population.

To satisfy its ambition to build a collaborative space to gather all kinds of data about migrant populations, the centre already has set up a data exchange scheme. The centre welcomes any institution holding data relevant to the migrant population that can contribute data to the collaborative space. And in return, the institution can directly obtain access to the data stewarded by the centre instead of applying through the annual window for access.

Currently, it is still at the early stage of establishing the data collaborative space. The question next is how the centre will go to build up new mechanisms and processes as well as build new technical infrastructure to support better coordination and collaboration among different types of data contributors. It is also worth exploring the potential risk of setting up such data collaboration to connect different kinds of data about the migrant population. For instance, whether it will unintentionally reveal personal privacy or whether it will allow private corporate organisations to identify a specific person with migrant status and refuse to provide certain services. It is critical to carefully evaluate use cases of data collaboration and set up novel rules to mitigate potential risks.

Shanghai Library

Shanghai Library stewards linked data and identifiers about names, locations, historic events and cultural works that are extracted and produced from Shanghai Library's rich collection of books, journals, photos, and multimedia. The library set up an open data platform in 2016 to make all data resources accessible to the public through open API[14] and SPARQL[15] endpoint under a Creative Commons licence. The library has been running an annual competition since 2016 to make data useful not only for professionals but also for ordinary citizens.

What data does the Shanghai Library steward?

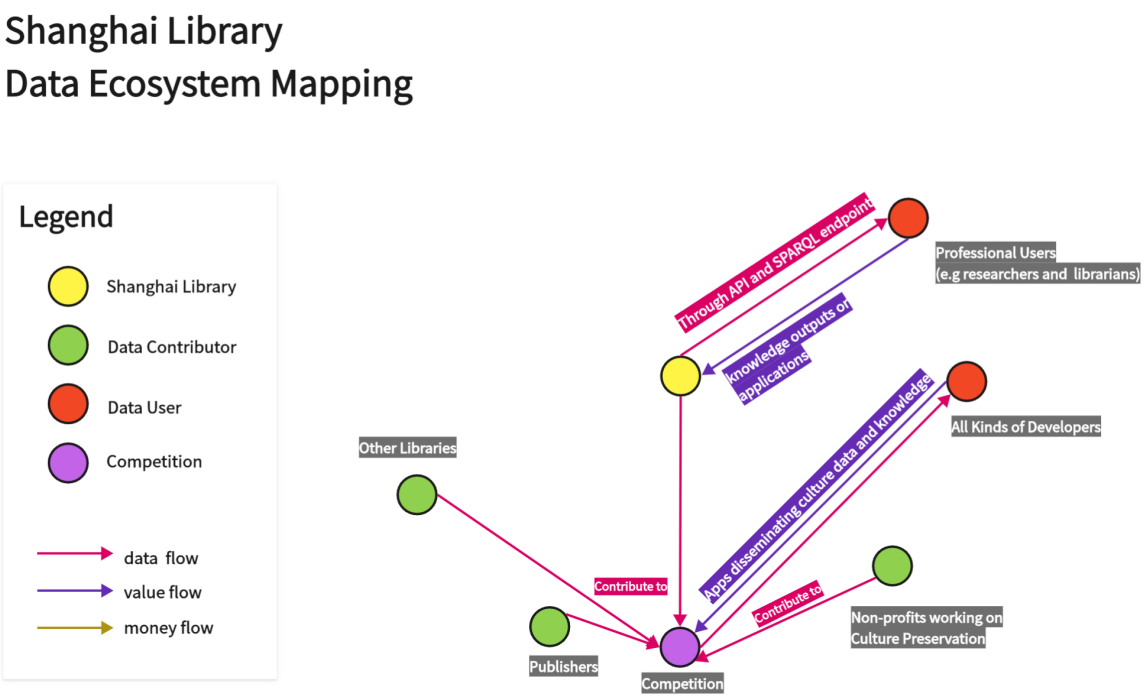

Figure 3: *Shanghai Library Data Ecosystem Map. Showing the relationship between the Shanghai Library (in the centre of the diagram) and: other libraries; publishers; competition; nonprofits working on culture preservation; developers; and professional users. The dataflows include: through API and SPARQL endpoint; and contributions. The value flows include knowledge outputs or applications; and apps disseminating culture data and knowledge.*

Shanghai Library extracted data from its rich collections to build structured databases and transform data into linked data. There are two categories of data that Shanghai Library has extracted and produced from its collections: general knowledge and literature knowledge.

The general knowledge base covers critical reference databases that serve as standardised identifiers. Those identifiers are harmonised names, locations, times,

[14] API is an application programming interface (API). It is a way for two or more computer programmes to communicate with each other.
[15] SPARQL is a query language used to retrieve and manipulate data stored in Resource Description Framework (RDF) format.

events, and objects. Currently, the general knowledge base includes data such as Chinese surnames (600+ records), geographic names (10,000 records), seals of historical organisations and famous figures (20,000+ records), as well as public institutions (20,000+ records). They were extracted from the literature held in the library and then harmonised and transformed into linked data.

The literature knowledge base transforms knowledge from existing collections of books and multimedia into linked data. The purpose is to turn literature knowledge into searchable and queryable knowledge graphs. It currently covers a wide range of content, such as family tree records, 'Shanghai Memories' photo collections, Shanghai local map catalogues, and old Chinese film records.

Since 2016, the Shanghai Library has run an annual competition to promote its data for wider use. There are interesting applications created by students, librarians, data scientists, and other types of users. For instance, one of the winning applications used the library's seal database to create a beautiful visualisation interface to browse Chinese ancient seals. Another example is to use the name database and metadata about books to create a game for users to learn about famous authors, their life events, and their novels.

Why make data accessible?

The motivation to build up the open data platform was driven by the desire to experiment with linked data techniques. The leadership at the library and the tech research team had a strong interest in making their own experiments on turning library data into linked data and making them available and accessible to the public was part of the experiment.

The library team was also inspired by what the Shanghai government was doing to build up the Shanghai Open Data Platform. As the open data culture started growing in Shanghai and later spread across China, the Shanghai Library felt more confident in making data open.

How to control access?

The Shanghai Library Open Data Platform requires users to register to obtain a unique token to further access data. There is no need for individuals or legal entities to apply for data.

The platform makes data accessible in three different ways: SPARQL Endpoint is the most powerful one that allows one to run complex semantic queries for example, 'who is the author of the TV show whose main character is called Doctor Who?'. The second way is the RESTful API provided by the platform which returns straightforward outputs as pre-designed by the API parameter (for example, an API call to "/get-author-name?bookname='The Tragedy of Macbeth'' will return "Shakespeare"). And the last one is through HTTP URI to visit the webpage listing the detailed structured data about the particular resource as pointed to by the URI.

There are currently no daily limits, but any high and frequent requests are being monitored. If any suspicious or unexpected high usage within a short duration is identified, the Shanghai Library team will be alerted and may contact the user to

further inspect and discuss a specific solution.

Highlights and reflections

Stewarding data infrastructure rather than just data

Data institutions tend to steward data, but in some cases, they also steward data infrastructure. Data infrastructure, as defined by ODI, is composed of data assets but also any people, processes, and technology that support the operation and maintenance of the data assets. Unlike the first case study, the Shanghai Library is a unique example of a data institution stewarding data infrastructure as well as data.

First, the data that the Shanghai Library stewards is not directly available as data but rather is extracted and produced from the rich collection of Shanghai Library. For instance, the Chinese name authority database is built on a rich collection of books and historical family tree records.

Second, it is important to note that the process of extracting information from books is not only to create name records but also to standardise and harmonise different expressions of names to create new identifiers. For instance, there are different ways to refer to the same person in China, because traditional Chinese names can be broken into three parts, including surname, first name, and style name, and they might be recorded differently across history and in books. For example, a person might be referred to by using only one part of the name or a combination of those parts in a different order. Thus, a unified identifier is created to refer to all those variations for the same person.

These two features make Shanghai Library's case unique. The Shanghai Library, as a traditional cultural institution, evolves into a data institution by digitising its culture collection and stewarding data. Furthermore, It also makes extra efforts to harmonise and standardise data and takes responsibility for stewarding common identifiers as infrastructure. Therefore Shanghai Library actually stewards not only data but also data infrastructure.

From a data institution to a network of data institutions

Since 2016, Shanghai Library has been running a competition to promote its open data resources and expects to engage wider communities to make use of its data. In addition to Shanghai Library data, the library also invites other data holders in the culture industry to join. As of 2022, the competition already has 18 different data partners ranging from similar libraries to publishers to other non-profit organisations in the culture sector, forming a network of data institutions stewarding similar sets of data and infrastructure.

The impact of the Shanghai Library Open Data Platform is thus not just about making data accessible and realising the value of data value. It also successfully promotes the culture of openness and serves as a tangible example for other similar organisations to start stewarding data and making data accessible. Currently, the network of data institutions is still at its early stage without too much coordination. It is possible to better coordinate those different data

institutions to better collect and process data as well as data infrastructure, to build a shared data common for the culture sector.

Summary and discussion

Revisiting the case studies

In this report, I identified two potential cases in China that fit the definition of data institutions. I conducted investigations to explore those data institutions with a focus on understanding their initial purpose and what kinds of measures they have put in place in order to facilitate safe access to the data they steward.

The table below summarises the key features of the case studies:

Data Institution	Data it stewards	Is data sensitive?	Channel through which the data is accessible	Access control measures
Migrant Population Service Center	Individual survey data filled by sampled migrants annually to report basic household information and attitudes to public service delivery	Yes. It contains individual household information such as wealth and health status	Through a public website: Migrant Population Data Platform	• Registration • Application • Agreement • Training • De-identified data • Direct download
Shanghai Library	Digital data extracted from or produced based on the collection of books, journals, photos, and multimedia	No	Through a public website: Shanghai Library Open Data Platform	• Registration • API and SPARQL Endpoint

Reflections on lessons learned

From data access programme to data institution

In this report, I studied two data institutions in China as part of a wider piece of work to study the concept in China. Both examples are existing institutions that are evolving into data institutions by taking new responsibility for data stewardship and setting up new data access programmes.

Take the Migrant Population Service Center as an example: it was originally established to support the National Health Commission to study and improve the public service delivery to migrant populations. It was later tasked to be responsible for stewarding the CMDS, the annual migrant population dynamic survey data. The scope of data stewardship at that time only covered managing the collection process and conducting data analysis to gain insights into data. The centre was later authorised to set up a data access programme in the form of a data-sharing platform website in 2014. And that is the moment that the centre started taking full responsibility of stewarding CMDS covering the whole data lifecycle – from collection to use to sharing.

But one thing I find interesting and tricky to answer is how to tell whether an institution is just simply a data access initiative or a true data institution? What is the line distinguishing these two? One comment I received via interview may help shed light on the future investigation:

'(There) are many platforms out there sharing and opening up data and they are often considered as a project or a programme within an organisation. But after years of operation, we find it may be a good idea to redefine who we are from the point of view of stewarding data and making data accessible as a service. It could be the right time to think about setting up a new institution dedicated to data stewardship'.

Building on this, I would suggest that one key feature of a data institution is making data stewardship the core purpose of the organisation. And to achieve this, in the examples covered in this research, it is necessary to establish a new organisation and transfer all duties of data stewardship from the existing organisation to the new one – to make it become a data institution. Moreover, a data institution seems to steward 'thick' data, meaning not only the data resource but also the data infrastructure including the relevant standards and identifiers. This was observed in the example of the Shanghai Library where it does not simply make its data accessible (like a data access initiative) but also stewards a set of identifiers such as the Chinese name identifier database.

Same goal but different expectations

Maximising the value of data is the ultimate and shared goal of the data institutions investigated as part of this research. Several organisations identified in the register have been tasked by their respective superiors to mine data value at the very beginning of their establishment, while others, like the Shanghai Library, have no such duty.

I observed that this small difference leads to different expectations of how external

users will use data: Shanghai Library welcomes any new and innovative use of its data even if it is completely irrelevant to its work as a library. In contrast, the Migrant Population Service Center has to make a balanced choice: they welcome novel use of their data that falls outside of their original scopes but also put heavy weight on finding external partners who can deliver relevant data insights to fulfil their original duty of business (improving public service delivery to the migrant populations).

I also noticed that none of the data institutions included in the China Data Institution register currently have any pressure on thinking about how to sustain their own operations. Part of the reason is most of them are currently publicly funded or part of giant tech companies. But some of them are starting to plan their own business models. For instance, one of the initiatives stewarding electronic vehicle data explored how to deliver useful data insights as a paid service to car and battery companies. And in turn, it would like to receive more data access applications from data users who can deliver such insights to develop a data service market that benefits the data institute itself but also data users and other stakeholders.

Competition as a strategic tool

Increasing the visibility of data resources that are accessible to the public is a critical issue for data institutions to address. One common approach employed by most organisations in the China Data Institutions register is to run competitions or similar events to boost visibility and attract potential users to make use of data. But beyond simply increasing visibility, competitions can also help data institutions to achieve other goals.

First, it is found that competition as a strategic tool can help data institutions to form different types of communities. Developers and data scientists are usually attracted and later engaged by data institutions to turn data into applications or insights. But in addition to the developer community, there are also other communities serving different purposes. The Migrant Population Service Center, for instance, leverages the competition process to form an expert community, which is composed of expert competition participants and also expert judges invited by the centre. This expert community then will be engaged not only during the process of the competition but also beyond the competition. It will advise the National Health Commission on migrant population policies by reporting lessons learned from local practices or sharing the latest academic research.

Second, competition can also be a good tool to efficiently collect and review a large number of applications within a short dedicated time window. In the case of the Migrant Population Service Center, it runs an annual competition and sets the competition time window as the only time the data institutions accept any external application. By doing so, the centre is able to save its workforce from processing a large number of applications randomly throughout the year.

Last but not least, competition also helps data institutions to build a network of data partners or data institutions. In the case of Shanghai Library, the library's open data competition helps attract a network of data partners to contribute data. Some of them are libraries too, but others are cultural institutions or publication companies. This potentially leads to the creation of a data institutions network where different data institutions within the same domain join together to offer data

access and steward common data infrastructure.

Future direction

In this report, I investigated two existing publicly-funded data institutions in China as part of a wider project to research data stewardship in China. I plan to conduct more case studies in the future to better understand diverse types of data institutions in China.

I also have not yet made any concrete comparison between data institutions in China and those outside of China. It would be interesting to make such a comparison once I have much richer collections of data institutions in China. Given the different cultures and social contexts, it would be interesting to explore how data institutions as a concept are understood and adopted differently and what kind of governance structure and business models different data institutions employ.

I am also interested in understanding how different data institutions in different countries that steward the same type of data can collaborate, as it is critical to the global digital trade and also to addressing global issues such as climate change.