PD MANAGEMENT CONSULTING LIMITED

Feasibility study for an international metadata platform
The Trusted Data Observatory (TDO)

Author

Pádraig Dalton, PD Management Consulting Limited
Consultant and former Director General of the Central
Statistics Office of Ireland

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Foreign Affairs FDFA

Federal Department of Home Affairs FDHA
Swiss Federal Statistical Office FSO

November 2024

Table of Contents

Executive Summary

The November 2014 report, entitled "A World That Counts: Mobilising the Data Revolution for sustainable Development" included a quote that captures the key role that data plays in the world in which we live – *"Data are the lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time; designing, monitoring and evaluating effective policies becomes almost impossible"*.

Not all data sources are equal and fake news and alternative facts unfortunately have become terms with which we are all too familiar. An inability to distinguish between reliable and unreliable data sources is a challenge for those who want, and indeed need, to make informed decisions and those that want to live in a society informed by robust, reliable and trusted information.

Data is the foundation for informed decision making but has little or no meaning without metadata. Metadata is "data about data" and is the key to data discovery, and an assessment of accessibility, and comparability. In many cases it is the absence of structured, standardised or indeed machine readable metadata that creates difficulties when trying to distinguish fake news/alternative facts from robust and reliable information.

This report explores the feasibility of the development of an international metadata platform, the *"Trusted Data Observatory"* (TDO), based on standardised, open, machine readable metadata to support Discoverability, Comparability and Accessibility of trusted data sources. The proposed observatory is a metadata platform as distinct from a data platform and there will be no data on the platform which remains within the data compilers' environment and control. The objective is to raise awareness and more specifically, discoverability, of trusted data sources especially for users who depend on third-party search engines, an increasing number of which are now AI enabled, to find/source their data/information.

The TDO will only support the discovery of "trusted" data sources and in the initial phases of the project the TDO will focus on the data holdings of National Statistical Institutes (NSI's) and International Statistical Organisation or International Organisations with dedicated statistical divisions (IO's). As the TDO evolves over time the inclusion of a broader suite of trusted data sources can be explored and incorporated.

The proposal for the development of the Trusted Data Observatory is unique in how it proposes to use a global metadata platform to support Discovery, Comparability, and Accessibility. Its development will not be without its challenges and a phased, evolutionary approach to development will be necessary.

The challenges include the absence of metadata or indeed metadata of varying quality, the usage of different metadata standards, the risk of creating a metadata swamp, identifying the appropriate governance structures for the observatory, the business model and how it can be sustained over time, the incentive for data compilers to develop their metadata and engage with the observatory, and creating an active user base for the observatory.

The potential added-value is significant. It is difficult sometimes to place a value on issues such as discovery and visibility of trusted data sources, but perhaps simply by looking at the very evident negative and detrimental impact of fake news and alternative facts on democracies and civil society, one can begin to see the value that the TDO can provide to user groups including researchers, the media, non-governmental organisations, policy makers and legislators, governments, academia etc. The TDO will also support improved search results/answers from AI enabled search engines, drive users to and usage of trusted data sources, and support the harmonisation/standardisation of open metadata. It also addresses the challenge of "metadata silo's" where the application of different standards across multiple data sources/platforms can complicate data discovery and comparability.

The crowded data market and the proliferation of fake news and alternative facts place a premium on the discovery and use of trusted data sources which is at the heart of this proposal and is grounded in the desire to support decision makers to make informed decisions based on trusted data and to provide an environment where the public have access to data sources/information that can support them to live in an informed society.

1   Background to feasibility Report

The Swiss Federal Department of Foreign Affairs in collaboration with the Federal Statistics Office of Switzerland are exploring the feasibility of creating a Global Metadata Platform to support data discoverability and transparency in a world that is in the midst of a data deluge. This feasibility report has been commissioned to explore the scope, functionality, challenges and opportunities associated with the establishment of an international metadata driven platform. The Key Research Questions for the project can be found in Appendix 1.

2   The problem statement

Not all data is good, and not all data is bad, but how do we distinguish the good from the bad? Fake news and alternative facts unfortunately have become terms with which we are all too familiar. Decision makers, those developing and evaluating policy, researchers, Non-Governmental Organisations (NGO's), the media, academics and indeed civil society increasingly need mechanisms to support them to discover and subsequently access robust, reliable, verified, coherent, comparable, high-quality and most importantly trusted data and information.

"Data are the lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time; designing, monitoring and evaluating effective policies becomes almost impossible."

This is a quote, from November 2014, taken from the United Nations Secretary-General's Independent Expert Advisory Group report on a Data Revolution for Sustainable Development, entitled "A World That Counts: Mobilising the Data Revolution for sustainable Development"[1]. This quote accurately captures the critical role that data plays in today's world.

Turning data into information and knowledge

DATA → INFORMATION → KNOWLEDGE → INFORMED DECISION-MAKING

creates — which generates — and leads to

The World Development Report 2021, Data for Better Lives[2] also highlights the critical role data plays in an ever-changing world and places a particular focus on data governance, a key part of which is metadata.

Data is the starting point for information, knowledge and insight. Finding the data you need when you need it, especially from trusted or Official data sources can be difficult. If you are an expert user, then finding reliable data can be relatively straight forward. However, beyond the community of experts, finding reliable/trusted data can be far more complicated. Many users are dependent on lexical search engines (e.g. Google, Bing), and AI enabled search engines (e.g. ChatGPT, Google Gemini) in this regard. These tools trawl data holdings in the virtual world and attempt to make sense of the information available to them and present that data in an understandable and credible manner.

The capacity and ability of these engines to explore every corner of the virtual world is a strength but also a weakness as these tools frequently identify data of mixed quality. The Office of the Chief Statistician of the World Bank has undertaken a considerable amount of work in this area and has highlighted the challenges that can arise when using third party tools of this type to discover trusted data. AI and LLM based search engines often select content from multiple sources and bring them together to provide a combined answer that can often be mis-leading as the information to identify the reliable from the unreliable sources is not always available.

Data has little or no meaning without metadata (see section 4 and Appendix 3) and in many cases it is the absence of structured, standardised or indeed machine readable metadata that creates difficulties when trying to distinguish fake news/alternative facts from robust and reliable information.

Clearly the absence of metadata and an inability to distinguish between reliable and unreliable data sources is a challenge for those who want to live in a society informed by robust, reliable and trusted information. This, it can be argued, creates an environment that supports the spread of misinformation[3] and disinformation[4] and by extension contribute to an undermining of civil society and democracy in many parts of the world today.

Fake news and alternative facts will, unfortunately, always be with us and it is important to recognise that the development of the proposed metadata platform will not solve this problem. The objective is to raise awareness and more specifically, discoverability, of trusted data sources especially for users who depend on third-party search engines, to find/source their data/information. Trusted data is getting lost in the "data swamp" and it is of societal importance, that at a minimum, we take steps to make trusted data sources discoverable and accessible. Standardised metadata, that is consistent with machine readable and AI enabled technologies, while avoiding the creation of "metadata swamps" and "dispersed metadata silo's", is an important first step in bringing trusted data to the fore and into the conversation.

There is no doubt that there are many platforms (primarily data as distinct from metadata platforms) in existence today that contain reliable and trusted data but these sources are getting lost in the "data swamp" that now exists across the virtual data world. This is a challenge that needs to be addressed.

---

1   See A-World-That-Counts.pdf (undatarevolution.org)

2   World Development Report 2022: FINANCE for an Equitable Recovery (worldbank.org)

3   Inadvertent spread of false information without intent to harm

4   Deliberately spread false information designed to mislead others and confuse fact and fiction

3   A Trusted Data Observatory –
Definition and Scope

The vision for the proposed "Trusted Data Observatory" (TDO) is based on the usage of standardised, open, AI enabled, machine readable metadata to support Discoverability, Comparability and Accessibility of trusted data sources.

The proposed observatory is a metadata platform as distinct from a data platform. The data will remain within the data compilers' environment, and control over who and when the data can be accessed remains within the sphere of competence of the data compiler. The observatory is designed only to support the discovery, comparability and accessibility of data through metadata. No data will reside on the observatory.

The obvious question is how to define or identify "trusted data sources" in the context of the TDO. In the initial iterations of the TDO it will be necessary to make some assumptions around what constitutes a trusted data source.

Initially, at least, the TDO should limit its scope to the data holdings of National Statistical Institutes (NSI's)[5] and the holdings of International Statistical Organisations or International Organisations with recognised statistical Divisions (IO's). Such organisations compile their statistical data within the framework of the UN Fundamental Principles of Official Statistics[6] and in line with internationally agreed statistical standards and classifications.

It is important to acknowledge that there are trusted data sources beyond NSI's and International Statistical Organisations or International Organisations with recognised statistical Divisions (IO's), and the inclusion of these sources is certainly something that will need to be explored as the TDO develops and evolves over time. Focussing on a phased or evolutionary approach suggests a starting point that is focussed and limited in scope, initially at least. The central point is that the TDO should be limited to supporting the discovery of trusted data sources. To do otherwise would defeat the purpose of the TDO.

Trusted data sources, like NSI's, compile a broad range of products across a very broad range of themes and the scope of the data holdings to be referenced through the TDO may include:

- Statistical releases and publications
- Aggregate indicators
- Online aggregate data bases where users can "build" their own tables
- Public use microdata files
- Synthetic microdata files

- Restricted access microdata files where the policy around access and the mechanisms to apply for access is transparent and obvious to the user. The role of the observatory should be limited to the provision of metadata to the user so that there is transparency around the content and potential value of the data. Access to the data would be outside the remit of the TDO but, as referenced above, the observatory should provide transparency around the access policy and mechanisms to apply for access. Access can only be facilitated by the data compiler in line with relevant policies and legislation

The audience or user community for the TDO is broad. The user community is focussed on those who need access to robust, reliable and verified data, or in other words trusted data. This community will incorporate the general public, the business community, legislators and policy makers, non-governmental organisations, the media/data journalists, the research community and academia. It should also include those involved in supporting others to find trusted data like many of the big technology companies, the Open AI community and the Open Source community. Data compilers may also be users of the observatory as well as contributors of metadata.

In developing the TDO it will be important to engage with as many of these user groups as possible to ensure their needs are reflected to the greatest extent possible. Doing so could play a key role in driving users to the TDO once it has been established.

The proposed observatory is a metadata platform. Addressing underlying existing data challenges, such as definitional differences or scope limitations, is not something that can be achieved directly by the metadata observatory, however the TDO could play a useful role in highlighting consistency and comparability challenges across data holdings which can be a starting point for addressing such issues.

---

5   Sometimes also called National Statistical Offices (NSOs)

6   See https://unstats.un.org/unsd/dnss/hb/E-fundamental%20principles_A4-WEB.pdf

4 The TDO Model – Input, Throughput and Output

The Director General of the Federal Statistical Office of Switzerland, Mr. Georges-Simon Ulrich, has articulated what he refers to as "The value chain of data products: input, throughput and output". In many ways his description of the input, throughput and output model provides a useful insight in supporting an understanding of what the Trusted Data Observatory is aiming to do, and is included here to provide additional insight:

In general, data products follow a clear value chain, which is divided into three main areas:

- Input: Where does the data come from?
- Throughput: Who can access the data and how?
- Output: Who makes products out of the data?

Input area: The challenges

The biggest problem in the input area is that we often don't know what data we have and how we can use it.

A common misconception is that text data represents all data. This is wrong. Data exists in many forms and it is crucial that it is recorded and organized correctly.

Data equity: Data is not neutral – it is the result of methodological decisions. The state should play an enabling role here and set a good example with representative administrative data.

Throughput area: Access to data

Platform economic giants such as Google, Alibaba and Amazon are successful because they control all three areas of the value chain.

The state should ensure that access to data remains fair and that no one is disadvantaged. The human being must always be at the centre of this.

One possible solution would be to create a **metadata platform** to make data visible and accessible. This could be made possible through legal, technical and organizational measures.

Start where such networks already exist, which is why ECOSOC[7] has commissioned the United Nations Statistical Commission, to work on issues such as highlighting the advantage of representative data, agreement on definitions along the entire value chain, including for topics that go beyond purely geocoded questions. Process of agreeing on objectives such as Sustainable Development Goals (SDGs), developing common frameworks (quality frameworks), national responsibilities, discourse in the UN.

Output area: creating data products

ChatGPT is an example of how data is used to create products. It accesses input data and genErates answers using statistical models. Similarly, in the future, information products of official statistics or generative AI applications could be created based on data.

5 Metadata – what is it?

As outlined earlier, metadata[8] is the key to data discovery and an assessment of accessibility, and comparability. In the world of Official Statistics, broadly speaking, there are two types of metadata: structural metadata and reference metadata:

- Structural metadata are used to identify statistical data such as titles, subtitles, short descriptions, dimension names or variable names etc.
- Reference metadata describe statistical concepts and methodologies used for the collection and generation of data. They provide information on data quality and, since they are strongly content-oriented, assist users in interpreting the data

Metadata standards are instrumental in improving interoperability in official statistics by establishing common formats, structures, and definitions for data and accompanying metadata. They facilitate the seamless exchange, integration, and understanding of data across different systems and organisations. Examples of metadata standards that are relevant to official statistics are set out in Appendix 3.

This report does not intend to explore the technical elements relating to metadata standards, which are complex in nature, but what is important is that the metadata on the TDO is structured and standardised to support machine readable and AI enabled technologies making the data more visible. This approach will also facilitate an API enabled observatory where metadata can be pushed to the TDO by the data compiler and pulled by the TDO from the data compiler.

One of the significant risks we must try and avoid in developing the TDO, is the creation of a "metadata swamp". If we try and replicate all of the detailed metadata that is attached to data at source (in the data compilers' environment) on the metadata observatory then we certainly run the risk of creating such a swamp.

There may be a benefit in identifying a "Minimum Viable set of Metadata (MVM)" that is designed solely to support the core objectives of data discovery, comparability and accessibility. This MVM could be identified in consultation with metadata experts, user groups and data compilers. This approach may limit the amount of work data compilers need to undertake to ensure their

---

7   Home | Economic and Social Council

8   See Overview – Eurostat (europa.eu)

metadata is structured according to the agreed standard (which ideally should be an existing standard) for the TDO. Users will still be in a position to access to full suite of metadata attached to any given source within the data compilers' environment, once the data has been discovered.

6   Functionality of the Trusted Data Observatory

The proposed scope of the metadata (and by extension the data sources) to be referenced through the TDO was discussed earlier (see Section 3). It may be prudent in the initial phases of the project, and particularly during the Proof of Concept phase, to limit the range of themes to be referenced by the observatory and focus on one or two key areas (e.g. SDG's. Labour Market).

The functionality of the TDO needs to be focussed on supporting the vision for the observatory and in particular the core objectives of Discoverability, Comparability and Accessibility of trusted data sources.

The functionality of the Trusted Data Observatory may include:

- The extraction and receipt, via API's of well structured, appropriately indexed and standardised metadata

- An advanced search functionality to support data discoverability. In addition to the traditional lexical search function (where a search engine looks for literal matches of the query words or variants of them, without understanding the overall meaning of the query) consideration should also be given to the inclusion of a semantic search function (which seeks to improve search accuracy by understanding the searcher's intent and the contextual meaning of terms as they appear in the searchable dataspace[9]).

- The ability to use appropriate "filters" to tailor or restrict searches, once on the observatory, based on specified categories including for example, data sources (e.g. NSI data only), themes (e.g. SDG's, agriculture, employment), frequency, periodicity etc.

- Potentially the use of AI tools to interrogate the metadata associated with a specified search to highlight data gaps, differences or comparability issues across the data sought in the search

- Provide a tracking function to identify user cohorts and key areas of interest

- Provide a "feedback" option to users to support the ongoing development and evolution of the observatory

- Provide a user friendly interface to promote and support the use of the platform

---

9   See Semantic search – Wikipedia

In setting out the functions for any new initiative of this nature the risk is always the temptation to "over specify" or "over reach". Taking a focussed approach and keeping the observatory as simple as possible from a functionality perspective, in the initial phases at least, may be beneficial. It is perhaps better to execute three or four elements correctly rather than attempt to execute a multitude of elements and not deliver fully on any of them.

7   The value proposition

Traditionally the focus of compiler's has been on the data, with the organising and structuring of the metadata seen as secondary in some way. That approach is now changing as the value of data, and by extension metadata, has been recognised. It is particularly important in the modern data ecosystem, given the number and diversity of data compilers now operating in the crowded data market and the broad diversity in the quality of data products and sources, that a clear focus is placed on not just having up-to-date, open and well-structured data, but also to have up-to-date, open and well-structured metadata.

The discovery of trusted data within the context of an increasingly crowded data market is a key challenge that needs to be addressed. The world we are now living in is complex as are the information needs of decision makers and others. Discovery of trusted data is an important first step.

Within that broad context the value proposition of the proposed platform may include:

- **Discovery and visibility:** Supporting the discovery, comparability and an assessment of the accessibility of trusted data sources through one observatory using standardised and structured metadata. This will provide a level of transparency and consistency to how data is discovered and what data is available to users. Providing this information, on one platform with standardised metadata, removes the proliferation challenge of multiple platforms with varying metadata standards (where metadata exists) and supports the increased visibility and discoverability of trusted data sources to third party search engines including AI enabled search engines

- **Improved search results/answers:** AI enabled search engines focus on providing "an answer" to a question, rather than a list of search results to the user, based on the data identified during the search. In many cases the answer provided to the user is the amalgamation of data extracted from various sources often of varying quality with limited metadata to support the user to assess the source or quality of the information provided. Bringing the metadata for trusted sources together on one platform certainly

increases the opportunity and likelihood of third party search engines discovering trusted data and providing users with more robust and reliable answers

- **Training algorithms:** Large Language Models (LLMs) need good quality data to train their algorithms. The availability of a platform with standardised metadata, pointing only to trusted data sources, if used appropriately, has the potential to support the improvement of the performance of LLM based tools for the benefit of all

- **Drive usage:** The TDO, if successful in implementation, will drive users towards data compiled by NSI's, IO's and other trusted compilers of data. Maximising the usage of Official Statistics is a core objective for most, if not all, compilers of Official Statistics, and this may act as a motivating factor for data compilers to engage with the project

- **Avoid metadata silo's:** The TDO places metadata in a pivotal position, and not just as an addendum to the data. The TDO will support the discovery and comparability of data and avoid the risk of the creation of "metadata silo's", using different metadata standards, which can work against the discovery of trusted and comparable data sources

- **Minimise compiler's work load:** Leveraging on existing metadata standards and limiting the amount of metadata to be provided to the TDO using the MVM approach, can minimise the amount of work data compilers/providers may need to undertake to meet the metadata requirements of the TDO

- **API enabled:** The TDO should be API enabled and in that sense the process of loading data to the TDO ("push" capability) would be automated. In addition, the TDO should also have a "pull" enabled API, which will also have the impact of minimising the workload of data compilers

- **Granularity and scope:** The range of data compilers, sources and products the observatory can link to if implemented, offers greater scope and potential for granularity in analysis

The value proposition can provide an impetus for trusted data compilers to develop and harmonise their metadata holdings in a manner consistent with machine readable and AI enabled technologies. The collection and storage of metadata in one place will make it easier to identify differences and inconsistencies in existing data and standards applied across countries and regions, and in that way may support the development of harmonisation and standardisation capacity building programmes over time.

The value proposition associated with the proposed observatory is significant, but for that added-value to be realised a significant effort will need to be undertaken to market and brand the TDO appropriately to create an active user base.

8   The Challenges

This is an ambitious project, especially in the context of its proposed global reach. Some of the main challenges to be considered include:

Challenges
Metadata – does it exist and if so is it harmonised/standardised
Risk of creating a "Metadata Swamp"
Governance of the TDO
Sustainability – the business model – who pays?
Incentive for data compilers' to engage with the TDO
"Scope creep" during development
Creating an active user base

These challenges are explored a little further here:

- The extent to which metadata is available and or harmonised/standardised. Even where it is available it may not be structured or adhere to a recognised international metadata standard. There has been an increased emphasis on metadata, especially within Official Statistical circles, but work remains to be done. The development of the TDO may be useful in helping to uncover the scale of the work yet to be undertaken and support the development of capacity building strategies in this field

- Avoiding the creation of a metadata swamp – see Section 5.

- Governance is certainly an issue that needs to be considered and addressed. The observatory is global in nature and in that context, it makes sense that it is governed at a global level. The ambition for the platform has come from Switzerland and both the Federal Statistics Office (FSO) of Switzerland and the Swiss Federal Department of Foreign Affairs are committed to turning the ambition into something real and tangible. That said, there is a clear recognition that ultimately the governance of this global observatory, even though it will be located in Geneva, must involve the global community. It may be useful to reflect on the role the community of Chief Statisticians of NSI's from around the globe can play via the United Nations Statistical Commission (UNSC), and also the role the Chief Statisticians of International Organisations can play via the Committee for the Coordination of Statistical Activities (CCSA)

- The "business model" underpinning the TDO needs to be considered so that the sustainability of the observatory from a funding perspective can be assessed. Data is a "Public Good" and in that context the TDO should also be seen as a Public Good

where usage of the observatory is free of charge. Creating hurdles or barriers to usage would be counter-productive and the objective must be to maximise discoverability of trusted data from trusted data sources. This leaves the question of funding and, by extension, sustainability open, and it should be anticipated that voluntary contributions or provision of in-kind supports from interested parties and those that recognise and benefit from the value-add that it provides will be required

- The incentive for data compilers to engage with the TDO is certainly an important consideration. Section 7 above provides a strong rationale for engagement and highlights the value-add that can accrue to data compilers to encourage their participation. It will be important that data compilers play a key role in the stakeholder engagement process that will need to be undertaken should a decision be taken to proceed with the project

- Being overly ambitious, at the outset, in the specification of the proposed TDO is a risk. Scope creep is always a risk in projects of this nature, and it will be important, especially for the Proof of Concept exercise, to be clear on the objectives and to limit the scope and function to that which is required

- Creating an active user base will be a challenge. There are many examples of good and useful platforms, but all too often they are underutilised. Engaging with users and data compilers in a meaningful way will be extremely beneficial if the project is to be successful. Involving marketing and branding experts will also be a necessary step and a clear marketing and branding strategy will need to be developed focussing on key user groups, the value-add, the benefit to users, the benefit to data compilers and of course the benefit to those supporting data discoverability through search engines, chat bots and LLMs

9 Next steps in creation and implementation of the proposed observatory

It will be important to use an agile methodology in the development of the observatory. The following provides a proposal for a suite of steps or phases of development towards full implementation of the TDO:

Phase 1
Identification of all relevant stakeholders and detailed engagement to identify needs and develop specification for the TDO
	To kick-start this process there may be a value in organising a "scrum" or "sprint" involving key players such as the World Bank, OECD/PARIS21, IMF, UNSD, UNECE, Eurostat, and a selection of NSI's.
	    This is a critical phase in the process. The skills required to support this activity may include project management, statistical expertise with knowledge of both data and metadata standards, and systems and business analyst skills.
	    Identification of seed capital and governance for PoC should also be identified.
		2025

Phase 2 Identify and engage with potential partners to participate in the Proof of Concept (PoC) exercise	The objective here should be to identify a suite of interested partners (a collaboration of the willing – NSI's and IO's) to participate in the PoC. These partners should reflect the global ambition of the project and perhaps also include NSI's with varying levels of maturity around metadata. This may be useful to assess potential workload for NSI's and how capacity building support can be offered within the framework of the project. Consider a role for universities/academics/philanthropy.	**2025**

Phase 3
Observatory development (prototype) for PoC
	Complete the technical specification for the observatory. Observatory goes online, core issues are resolved, and information is accessible. Develop prototype in a manner that is scalable. Encourage the inclusion of the observatory in UN strategies. Switzerland to lead with partners
	    Ensure sustainable funding & governance structures.
	    Commence work on an engagement and communications plan to generate knowledge and interest in the platform.
		By end 2026

Phase 4 Review of PoC	Review and refine the TDO post PoC exercise	**By end 2027**

Phase 5
Onboarding & "Post-2030 Agenda"
	Onboard key partners (NSI's and IO's not involved in PoC) and integrate with the "Post-2030 Agenda".
	    Build on communications and engagement plan from phase 3 to drive engagement of data compilers with the observatory and drive usage of the observatory by the user community.
	    Examine a broader suite of trusted data sources within an agreed framework developed by the governance structure put in place for the TDO.
		2027 – 2030

10 Geneva as the location for the platform

We live in a world where data can move, pretty seamlessly from one part of the world to another, almost instantly. In that sense a metadata platform could in practice reside in any corner of the globe.

That being said there are very good reasons to consider Switzerland, and in particular Geneva, as an ideal location for the proposed Trusted Data Observatory.

Some of the key considerations are as follows:

- Geneva acts in many ways as a hub for many global organisations, including "data-heavy" organisations, and as a result the infrastructure, the global reach and the experience of operating at this level is very evident and present in Geneva. See Home – GIPLAT-FORM which is "The Geneva Internet Platform (GIP)", initiated by the Federal Department of Foreign Affairs (FDFA) and the Federal Office of Communications (OFCOM) of Switzerland in 2014, that provides a neutral and inclusive space for digital policy debates, recognised by the majority of global actors as a platform where different views can be voiced, and The UN in Geneva | The United Nations Office at Geneva (ungeneva.org) for a list of UN organisations located in Geneva

- Switzerland is a neutral country and, in that context, locating the platform in Geneva can be globally accepted

- The ambition, leadership and vision to develop such a platform has originated in Switzerland and there is a real commitment to support the development and implementation of the observatory

- Switzerland has been home to international organizations for over 130 years. Throughout this long-standing tradition, Geneva has developed a unique infrastructure specifically designed to meet the needs of international organizations. Swiss and Geneva authorities are highly responsive to these needs and continuously work to enhance Geneva's appeal as a host city for such organizations. Today, over 40 international organizations and more than 170 international non-governmental organizations have their headquarters in the lake Geneva region. In addition, elements of the technical infrastructure required to host the observatory may already be in existence in Geneva e.g. the United Nations International Computing Centre (UN ICC) could be a valuable partner in this project

11 Examples of other platforms

There are few examples of metadata platforms of the type envisaged for the Trusted Data Observatory. In many cases metadata is associated with the data where the metadata is accessed by first of all gaining access to the data. The process flow is Data to Metadata, one data source to the associated metadata. This is the model used by many National Statistical Institutes and International Statistical Organisations.

The approach proposed for the Trusted Data Observatory has the metadata as the starting point where the metadata is used to discover the data. This approach facilitates users to identify multiple data sources for the prescribed metadata and in that sense one metadata source is being used to identify multiple relevant data sources from contributing data compilers.

Many of the examples found could be described as Data Catalogues, very often where the data resides on the platform. However, there are some nationally focussed metadata platforms referenced below, but with limited links to the data. The list below is not exhaustive and is provided simply to provide a sample of what is already in existence.

- SSBs Metadata – Statistics Norway's metadata systems

- I14Y Interoperability platform – The I14Y interoperability platform is Switzerland's national data catalogue. It ensures the efficient exchange of data between authorities, companies and citizens. In the platform, an overview of the data collections and interfaces of the Confederation, cantons and communes is continuously expanded and their metadata are made available centrally

- Insee – The Melodi application (My Extract and Load Open Data at INSEE) offers statistical public datasets in open formats

- RMéS, INSEE's statistical metadata repository – N2 Statistics Courier – 2019 | Insee

- https://integrateddataservice.gov.uk/data – The Integrated Data Service (IDS) is a cross-government project, with the Office for National Statistics (ONS) leading its delivery. The IDS is a central platform that provides access to de-identified data. You must be an accredited researcher to access the full IDS data catalogue

- https://www.ine.es/en/HVD/ – This is a Data Catalogue and it provides access to the High-Value Datasets available for category 4: Statistics for which the National Statistics Institute of Spain is responsible

- https://www.adruk.org/data-access/data-catalogue/ – This is a Data Catalogue where the focus is the research community, academics and policy makers. The scope of data holdings relates to administrative datasets available for public-good research

- https://datacatalog.worldbank.org/home – The Data Catalogue is designed to make World Bank's development data easy to find, download, use, and share. It includes data from the World Bank's microdata, finances and energy data platforms, as well as datasets from the open data catalogue

- https://microdata.statistics.gov.rw/index.php/catalog – This is a Data Catalogue from the National Institute of Statistics of Rwanda

- https://catalog.data.gov/dataset/ – United States Government's open data site

- https://data.ontario.ca/ – A Data Catalogue from the Government of Ontario

- Integrated Data Infrastructure | Stats NZ – New Zealand – The Integrated Data Infrastructure (IDI) is a large research database. It holds de-identified microdata about people and households

- Open Government | Open Government – Government of Canada – Canada – Access to information that is disclosed as part of the federal government's commitment to enhance accountability and transparency

- Central Data Catalog (dane.gov.co) – Colombia – Central Data Catalogue

These examples, and ongoing developments across the globe, as compilers begin to standardise and harmonise around existing metadata standards and create national platforms, may be of benefit to the proposed Trusted Data Observatory.

12 Estimate of costs

The specification of costs before the full suite of requirements for the observatory have been outlined/specified are difficult to provide. For any project of this nature there will be initial set-up costs including a technology component and ongoing maintenance costs to support the technical infrastructure and support staff to maintain the observatory (content) and engage with data compilers and other stakeholders, including the user community.

The following tables provides an estimate of resources required for each of the phases outlined in section 9 of this report.

Phase 1	One team leader plus one/two IT specialist (systems/business analyst), and 2 scientific staff.	2025
Phase 2	The team leader outlined above should take the lead on this activity, therefore no additional cost to that outlined in Phase 1 above.	2025
Phase 3	One team leader plus four IT specialist (systems/business analyst), and 3 scientific staff. • Additional IT specialists (two) in support of phase 1 team will be required on a temporary basis to support the development of the prototype. • Additional scientific staff (one) in support of phase 1 team will also be required to support the population of the Observatory with metadata from the participating partners in the PoC. Technology costs relating to the hosting of the platform will need to be considered at this time. Costs will be dependent on market conditions. A Request for Information (RFI) and a subsequent Request for Tender (RFT) will be required to confirm costs. A marketing and communications consultant will need to be retained. Costs will be dependent on market conditions	By end 2026
Phase 4	No additional cost, use existing team members	
Phase 5	One team leader, 3/4 scientific specialists plus ongoing IT costs for hosting and incremental development of the Observatory as needed. A marketing and communications consultant will need to be retained. Costs will be dependent on market conditions	2027 – 2030

13 Conclusion

The proposal for the creation of a Trusted Data Observatory is ambitious and rather unique in how it proposes to use a global metadata platform to support Data Discovery, Comparability, and Accessibility. The rationale for the proposal is grounded in the desire to support decision makers to make informed decisions based on trusted data and to provide an environment where the public have access to information that can support them to live in an informed environment.

The crowded data market and the proliferation of fake news and alternative facts place a premium on the discovery and use of trusted data sources which is at the heart of this proposal. Creating a global metadata repository based on agreed metadata standards (machine readable) for reliable data sources with proper governance can help data users and search engines identify trusted data sources.

Engagement with data compilers and users of the new observatory will be essential during the development phase and beyond. Rome wasn't built in a day, and neither will the Trusted Data Observatory. This is a multi-year project, and a phased (agile) approach to develop and implement the observatory is set out in this report. The project will require strong leadership, a clear vision and a willingness for all to collaborate and engage at the global level. Establishing a collaboration of the willing will be an important first step in the project.

There is a clear recognition that the development of the TDO is just one of the initiatives that the broader data community will need to take to raise awareness and usage of trusted data sources. Capacity building and focussing on issues such as statistical and data literacy are examples of complimentary initiatives that will also be required over time. There is "no silver bullet" but the development of the TDO is a key initiative in supporting the discovery, comparability and accessibility of trusted data sources.

Appendix 1 – Key Research Questions

1. Assess the appropriateness and coherence of the project vision in relation to its overall objectives;

2. Assess potential value added for data owners;

3. Identify relevant examples of already existing meta-data platforms across regions and countries and which projects are currently under development;

4. Outline potential functionalities of the platform;

5. Scope of the platform;

6. Suggest criteria for the selection of (meta-)data to be referenced on the platform;

7. Identify realistic phases for the creation and implementation of the platform and suggest an approximate timeline;

8. Identify potential individuals to be involved in a pilot phase and in later stages of implementation, and their respective roles;

9. Assess the suitability of Geneva as the location for the platform;

10. Identify opportunities and risks in the different phases of the project;

11. Estimate which different cost factors (incl. staff requirements) there would be in developing (including the different steps) and maintaining the platform, including cost estimates where possible.

Appendix 2 – Stakeholder Meetings conducted as part of feasibility report preparation

- Tiina Luige, Director, UNECE Statistical Division
- Anu Peltola, Chief Statistician, UNCTAD
- Yannick Heiniger, Programme Coordinator Officer Giga, Jeremy Barnes, Senior Evaluation Planning & Reporting Officer, and Adeleh Mojtahed Senior Coordinator for Digital Inclusion Giga, International Telecommunication Union (ITU)
- David Jensen, Coordinator of the Digital Transformation Task Force, UNEP
- Benjamin Rothen, Ambassador, Head of International and National Affairs, Federal Statistics Office (FSO) Switzerland
- I14Y Team, FSO Switzerland
- Petra Keller Gueguen, Head of Staff Division, FSO Switzerland
- Georges-Simon Ulrich, Director General, FSO Switzerland
- Mathias Steffen, Project leader National Data Management, FSO Switzerland
- Bertrand Loison, Vice director and head of the division Data Science, AI, and Statistical Methods, FSO Switzerland
- Nina Frey, Deputy Head of the Digitalization Division, Federal Department of Foreign Affairs Switzerland
- Jean-Luc Bernasconi, Head of Analysis and Research, Swiss Agency for Development and Cooperation (SDC)
- Cristina Verones, Head of section UNGA-ECOSOC-HRC, Federal Department of Foreign Affairs Switzerland
- Francesca Giardina, Foundation Mercator Switzerland
- Peter Messerli and Andreas Heinimann, Wyss Academy for Nature
- Urmet Lee, Director General, Statistics Estonia
- Olivier Dupriez, Deputy Chief Statistician, World Bank
- Ivan Murenzi, Director General, National Institute of Statistics Rwanda (NISR)
- Steve MacFeely, Chief Statistician, OECD
- Johannes Jutting, and Francois Fonteneau, PARIS21

Appendix 3 – Metadata Standards

Examples of metadata standards that are relevant to official statistics include:

- SDMX: SDMX standard is a set of international standards for the exchange of statistical data and metadata. It is used by NSOs around the world to publish and exchange data
- DCAT: DCAT is a standard for describing datasets. It is often used to describe NSO data because it is designed specifically for datasets
- DDI: is an international standard for describing surveys, questionnaires, statistical data files, and social sciences study-level information
- GSIM: the Generic Statistical Information Model is a reference framework for describing the information objects that are used in the production of official statistics. GSIM is often used as a conceptual model including all Metadata needed to describe the statistical processes

The following links provide useful reference points to gain a more in-depth understanding of metadata and metadata standards and tools in use or in development.

Eurostat Reference Documentation

Overview – Eurostat (europa.eu)

World Bank Developments

- The metadata standards either currently in use or in the course of planning for implementation include the following:
  - For microdata: DDI Codebook 2.n
  - For geographic datasets: ISO 19139/19115/19110
  - For indicators/time series, the World Bank created a metadata schema by gathering elements from various databases (UN, OECD, WB, IMF, etc.). This schema will complement SDMX by offering a more structured and comprehensive framework for what should be included in SDMX's MSDs.
  - For documents/other materials related to datasets: Dublin Core + elements extracted from MARC21 (US Library of Congress)
  - For documenting research projects and scripts to enhance code discoverability, an internally developed custom metadata schema is used. This schema features in the new catalogue of reproducible WB research (https://reproducibility.worldbank.org).

- The World Bank is working on aligning these specialized metadata standards with the broader DCAT and schema.org frameworks to enhance SEO and GEO functionalities.
- All standards are defined in JSON format.

To promote the use of these metadata standards, the World Bank is creating several tools including:

- A Metadata Editor application intended to replace the Nesstar Publisher software and support multiple standards. The application is already used internally at the Bank. The application is complemented by an R package and Python Library (for advanced users who may want to exploit the application of APIs to automate tasks)
- AI-enabled tools for enhancing metadata (metadata augmentation to foster data discoverability and compatibility with generative AI applications)
- Cataloguing software adhering to these standards
- Advanced data discovery solutions aimed at evolving basic search (lexical) into sophisticated data recommender and knowledge retrieval systems (lexical + semantic)

Standards, guidelines, and related training materials will be openly accessible. All tools will be API-based and also openly accessible. One of the World Bank's goals is to enhance interoperability in data management and dissemination systems (enabling "federated data catalogues").

UNECE Common Metadata Framework

The Common Metadata Framework – High-Level Group for the Modernisation of Official Statistics – UNECE Statswiki

UNECE HLG-MOS Data Governance Framework for Statistical Interoperability (DAFI)

https://unece.org/sites/default/files/2023-11/HLG2023%20DAFI%20Draft%20Nov15.pdf