PD MANAGEMENT CONSULTING LIMITED

Trusted Data Observatory (TDO)

Phase 1 – complete
Phase 2 – contracted components

Full Report

trusted data observatory

Author

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

Commissioned by the

Swiss Federal Department of Foreign Affairs
Federal Statistical Office of Switzerland and
The PARIS21 Foundation Data For Change

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

April 2026

Table of Contents

Executive Summary 4

1 Introduction 6

2 Stakeholder mapping and engagement 6

2.1 Kick-off phase 6

2.2 Stakeholder engagement 7

2.3 Stakeholder survey and summary of findings 7

3 Governance structure for the TDO project 8

4 Risk Register for the TDO project 8

5 TDO Metadata standard(s) and MVM composition 8

6 Proposal for a Proof of Concept 9

7 Institutionalisation, governance, sustainability and hosting 10

8 Alignment with other AI related initiatives 10

9 Conclusions 11

Appendix 1 – TDO Stakeholder mapping exercise 12

Appendix 2 – Stakeholder Questionnaire 15

Appendix 3 – Frequently Asked Questions (FAQs) Document 23

Appendix 4 – TDO Participation Lists 26

Appendix 5 – Proof of Concept proposal for the TDO 27

Appendix 6 – TDO Business Specification 31

Executive Summary

The overall focus of the TDO project is on the development of an international metadata platform based on standardised, open, machine readable and AI enabled metadata to support Discoverability, and by extension Comparability, Accessibility and re-use of trusted data sources, for machines as well as persons.

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 Offices (NSOs) and International Statistical Organisation or International Organisations with dedicated statistical divisions (IOs). As the TDO evolves over time the inclusion of a broader suite of trusted data sources can be explored and incorporated.

The core elements of Phase 1 and 2 included:

- Stakeholder mapping and engagement
- Establishment of an appropriate governance structure for the TDO project
- Development of a risk register for the TDO project
- TDO Metadata standard and MVM composition identification
- Business specification for the TDO
- Proof of Concept exercise specification
- Considerations around institutionalisation and hosting of the TDO

Details of the various elements are set out in the main body of the report. The engagement with stakeholders has been extremely rich and insightful and is playing an important role in shaping the future of the project. Some of the key learnings from phase 1 & 2 of the project are summarised below:

1) Without exception the engagements with stakeholders have been extremely positive with every organisation met supportive of the TDO and willing to support the project in whatever way they can. This is a positive position for the project to be in at this point, but of course the project is moving into a new phase shortly, and what has brought the project to this point, may not be what is needed to bring the project forward beyond this point

2) The TDO initiative needs to have a very clear and influential strategic basis. This needs to be explored perhaps in the context of the International Geneva policy, or the UN80 Initiative work that will be brought to the 81st General Assembly in autumn 2026

3) Institutionalisation: These are perhaps the most important next steps for the project. Ownership and leadership needs to be invested in an organisation, ideally a globally focussed organisation. That leadership must be visible, committed and trusted.

4) Incentives to engage partners including LLM developers need to be considered

5) Trade-off: In terms of engaging the developers of LLMs, do they need to adapt to our environment, or do we need to adapt to theirs? The reality is that for the TDO to work, there will need to be movement on both sides (LLM developers prioritising the TDO as a point of entry to trusted data, and the TDO embracing Model Context Protocol, Generative Engine Optimisation etc). The Public Good and Civic Society perspective needs to be a central consideration for all parties in terms of their engagement

6) We also need to be clear on the value the TDO brings. Responses to the stakeholder questionnaire were extremely insightful in this regard:

- Observatory would make trusted data (e.g. official statistics) more visible, discoverable, and used internationally ensuring that national data remain relevant in global discussions and decision-making. By increasing visibility of trusted data through harmonised metadata standards, it strengthens confidence in trusted data sources, supports evidence-based policymaking and research and supports the pushback against misinformation and disinformation. It creates a central hub to find trusted data and provides an opportunity to meet the FAIR data principles[1]
- TDO provides a mechanism to reach agreement and convergence on the way forward with important strategic questions. A collaborative and cross-sectoral approach in pooling expertise from different organizations helps strengthen data governance and ensure diverse perspectives on standards, ethics, and methodologies. The initiative encourages cooperation between trusted data compilers, AI developers and data redistributors—an important step in ensuring that international standards are understood and applied beyond the traditional statistical community
- TDO would strengthen data governance and coherence across the community of trusted data compilers, by promoting harmonized metadata standards for data discovery (MVM), improving coordination between data producers, and supporting modernization and transparency in how data are documented and shared

1 https://www.go-fair.org/fair-principles

- Swiss neutrality is a valued asset and hosting the TDO in Switzerland is a positive approach and recognises the central importance of Data Sovereignty. Combining this approach with a global governance approach and institutionalisation of the TDO within a UN body highlights the global nature and ambition of the TDO
- Maximising the value and benefits of metadata is a significant benefit. By creating the TDO and "linking" what could be described as "metadata silo's", the benefits of metadata beyond understanding data can be realised, to include data visibility and discoverability
- The TDO can create a structured dialogue between private sector AI companies and the trusted data compiler community and help the official community "speak with one voice". An initiative that provides trusted data in a standard way for the whole community should be attractive for large tech companies to integrate with
- The platform's emphasis on machine-readability and open AI enabled standardised metadata could further enhance AI-assisted data integration and modelling
- Respect for Data Ownership and Access Protocols in the design of the TDO builds trust across all communities
- Global Collaboration and Capacity Building - TDO fosters international cooperation and could support countries with limited metadata infrastructure through shared standards and technical guidance
- The opportunity to replicate the TDO at national level as the TDO Open Source Software (OSS) code will be available for use in line with Digital Public Good principle

7) Governance is an important issue and is linked to "institutionalisation" highlighted above. The governance must reflect the global ambition of the TDO.

8) Sustainability: There is a hope that in time, as the value of the TDO becomes apparent, that in-kind funding can and will contribute to the sustainability of the TDO. This is a reasonable assumption and expectation when combined with dedicated and sustained funding from a relatively small number of funders. The costs of sustaining the TDO will not be prohibitive. The opportunities of co-development engineering focussed projects where specialists from across a number of organisations (multilateral approach) come together to undertake some technical work is worth exploring

9) Alignment: It is important to align/leverage other AI related initiatives already underway

Reflecting on progress to date, it is clear that awareness of the TDO is growing along with an clearer understanding of the role and value it can play in supporting the discovery of trusted data sources in an AI enabled world.

The next steps will be critical to the success of the project. **The institutionalisation of the TDO in a globally focussed organisation is the critical next step.** Once the institutionalisation issue has been addressed the key issues to address are as follows:

- Evolution in the governance of the TDO to reflect the global ambition and nature of the TDO project
- Agreeing on a host organisation for the TDO in Geneva
- Building the TDO application/platform in preparation for the Proof of Concept exercise
- Testing of the TDO platform before the PoC
- Establishing a PoC Participation Board to ensure the smooth running of the PoC
- Engaging with participating organisations in the PoC to ensure their metadata for the POC are aligned with the MVM and metadata standard(s)
- Conduct the PoC and evaluate the outcomes

Of course underpinning much of the above is the issue of sustainability. That is an issue that cuts across the whole project. To be successful the project will need a significant level of in-kind funding for participating organisations. The level of in-kind funding provided will have a direct impact on the level of monetary funding ultimately required.

1 Introduction

This report follows on from the previous work undertaken in 2024 relating to the delivery of a "Feasibility study for an international metadata platform – the Trusted Data Observatory". That report set out the challenge facing trusted data compilers in supporting the discovery of their trusted data in an AI enabled world.

Despite the fact that the Official Data and Statistics community already compiles high-quality trusted data with associated metadata, simple searches using a standard search tool, or indeed AI supported search tools, often provide plausible, yet inaccurate results. In this context there is a requirement to reflect on how we make our trusted data sources more visible and discoverable. A paper prepared by the World Bank and presented at the 56th UNSC in March 2025 entitled "Fostering AI-Readiness and Responsible Redistribution of Official Statistics" sets out the challenges very clearly.

It is important to note that trusted data compliers have high-quality metadata residing within their environments, but it is not always visible to traditional or indeed AI enabled search engines. The metadata standards in use are not tailored to the new AI enabled search/discovery environment and this is a fundamental challenge compilers of trusted data face. Without the use of harmonized metadata standards optimised for information discovery on the internet, even high-quality data can remain hidden from AI systems, reducing its discoverability and usability.

The overall focus is on the development of an international metadata platform, the *"Trusted Data Observatory"* (TDO)[2], based on standardised, open, machine readable (AI enabled) metadata to support Discoverability, Comparability, Accessibility and re-use of trusted data sources, for machines as well as persons.

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 Offices (NSOs) and International Statistical Organisation or International Organisations with dedicated statistical divisions (IOs). As the TDO evolves over time the inclusion of a broader suite of trusted data sources can be explored and incorporated.

This project is designed to support the push back against misinformation and disinformation by making trusted data (facts) far more visible and discoverable and is grounded in the desire to fortify the global data infrastructure which underpins democratic and civil society. It will not solve the problem of misinformation or disinformation unfortunately, but what it can do is support the process of putting trusted facts and figures in the hands of those seeking robust information (e.g. policy makers, NGO's, the media, civil society) thus providing a mechanism to challenge those trying to undermine democracy and civil society by spreading falsehoods, either inadvertently or purposely.

[2] See Trusted Data Observatory

The core elements of Phase 1 included:

- Stakeholder mapping and engagement
- Establishment of an appropriate governance structure for the TDO project
- Development of a risk register for the TDO project
- Final Phase 1 report to support subsequent phases

2 Stakeholder mapping and engagement

2.1 Kick-off phase

Stakeholder engagement by definition is an ongoing activity for any project that is designed to deliver a service to a defined user community. The kick-off phase in October/November 2025 provided some useful opportunities to engage on a face-to-face basis with key stakeholders. The event on the periphery of the celebration of the 20th anniversary of the European Statistics System Code of Practice was particularly beneficial as it brough key data compiler's together in one place.

Some important messages from the Kick-off/early stages of Phase 1 included:

- Reflecting on the fact that the issue of Data Discovery is complex, particularly when you start looking at the possible solutions. Communicating clearly is key and the language we use matters. Ensuring a common understanding of the problem and coalescing around a solution is undoubtedly a challenge. Clarity and transparency are important
- Elevating the role of metadata within the community of trusted data compilers needs work. Data compilers focus historically has been on the data, but now in an AI enabled world, metadata must receive, at a minimum, the same level of attention
- Experts don't always agree! Identifying an AI enabled metadata standard is an important next step in tandem with the identification of the metadata components that will constitute the Minimum Viable Metadata set. Agreeing on the MVM and AI enabled metadata standard may be a challenge
- While focussing on trusted data compilers in the early phase(s) of the project is important, it is equally important not to forget the broader community of stakeholders – the users and Big Tech
- Identifying the appropriate motivation to engage Big Tech will be challenging. The community of data compilers will need to come together with one (more powerful and influential) voice to maximise the potential to engage these important partners
- Strong and committed leadership of the project is a critical success factor

2.2 Stakeholder engagement

The compilers of trusted data are a key stakeholder group in the current phases of the project. Building their understanding and interest in the TDO is a critical step and will require ongoing attention. Much of the initial work is centred around some of the more technical data related issues (metadata) and in that context the compilers are a key group at this point. That being said it is also important to consider the potential user community and Big Tech, and in that context the Geneva Data Community and the Geneva Chamber of Commerce event have been useful.

A stakeholder mapping exercise was conducted to underpin key Phase 1 activities, and a copy of this report can be found in Appendix 1.

Our overall strategy from a stakeholder perspective involved:

- Developing key messaging around the TDO including the Vision, the burning platform (the Why), the longer-term strategy for the TDO and the value-add
- Identifying the suite of key stakeholders
- Identification of possible use cases for the TDO
- Providing clarity on the project plan
- Developing a stakeholder map
- Direct engagement with stakeholders (face-to-face, virtual, web and email communication)
- Undertaking a stakeholder survey focussing on data compilers, users and Big Tech organisations
- Identification of possible champions for the TDO and engaging directly with them

The engagement work continues and there is a clear recognition of the need to ensure that the project addresses the risk of being seen as an Official Statistics project. This evolution will commence in the latter part of phase 2 where governance structures and engagement plans will evolve to involve a broader community beyond compilers of trusted data. Emphasising the fact that the community of trusted data compilers goes way beyond the Official Statistics Community is also an import message for all users to hear.

2.3 Stakeholder survey and summary of findings

This section provides a brief summary of the feedback received from National Statistical Offices (NSOs), and International Statistical Organisations and International Organisations with a dedicated statistical division (IOs) stakeholders on the proposed Trusted Data Observatory. The questionnaire used to support this analysis can be found in Appendix 2.

The responses, 40 received in total, were extremely rich and provided valuable feedback on the proposed project. Some of the issues raised arose out of perhaps a lack of clarity in communicating the TDO and what it entails, and in that context alone the exercise was very useful. A Frequently Asked Questions (FAQ) document has been developed to address some of these communications issues and is attached as to this report (see Appendix 3) to support the development of a shared understanding of the TDO. The FAQ will remain a live document as new issues and questions arise. Indeed for those looking to gain a greater insight into the TDO, the FAQ is perhaps a useful place to start.

As outlined above the feedback provided was very rich and some of the key messages are summarised here:

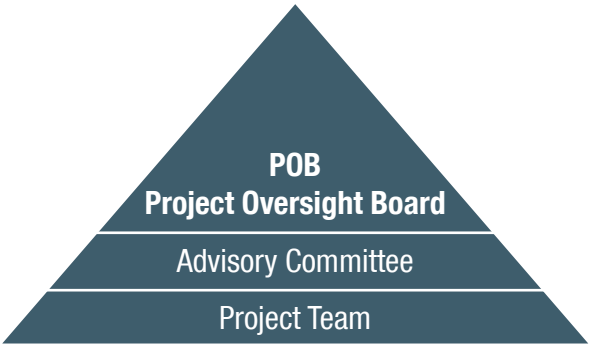
- There was a lot of positive feedback in relation to the vision and objectives underpinning the TDO. There is a clear recognition of the need to address the issue of Data Discovery in the context of an increasingly AI enabled world
- The need to align the TDO project with the broader suite of AI readiness projects was highlighted by many. The risk of duplication and the resultant need to split scarce resources across multiple initiatives was seen as a key issue to avoid. In particular the need for a multilateral approach was emphasised and the positioning of the TDO within this broader environment was seen as an important next step
- Similar to the point above, "Institutional Anchoring" was an issue that emerged from responses. It was emphasised that the TDO should be embedded within a recognized international statistical framework or hosted by a multilateral organization with a mandate in data governance. It was argued this would provide legitimacy, continuity, and alignment with global standards
- The metadata domain is complex. It was noted that standardising classifications across organisations/countries can be extremely difficult and could stall the TDO initiative. The number of internationally agreed standards in use across the community of trusted data compilers is considerable. Many respondent's recognised the ambitious nature of the project and in general there was a broad welcome for the Minimum Viable Metadata (MVM) set, as long as it did not result in the development of yet another metadata standard and did not require a significant amount of resources for compiler's to engage with and prepare the MVM. Fourteen respondent's expressed a willingness to participate in the metadata technical sub-group that will work on the identification of the MVM set, and the identification of the existing metadata standard to be used for the MVM
- The long-term governance and sustainability of the project was an issue highlighted by pretty much all respondents. Where will the funding for the TDO come from, can it really rely on in-kind contributions, who will decide what a "trusted" data source is etc. are just a couple of examples of the issues raised

- The incentives required to get both data compilers and in particular data redistributors (big tech companies) to engage constructively with the project was seen as a key area of work. Many emphasised the potential benefit of the multilateral approach and having the trusted data community speak with one authoritative voice
- Some issues raised in the feedback, went beyond the intended scope of the TDO. From a data perspective the issues raised are extremely relevant and include topics such as quality, definitional differences, variations in code lists used for variables etc. However, it should be noted that the trusted data community has been grappling with many of these issues for years and many different initiatives, including legislative change, have been employed to address these challenges, yet many remain. If we attempt to use the TDO project to resolve not just data discovery issues but the broader suite of underlying data issues, then the TDO project will not succeed. The value that can be attributed to the TDO project is very much aligned to the value put on "data discovery", which is the primary focus of the project

The questionnaire asked respondents to indicate their willingness to participate in the Proof of Concept and the Metadata technical sub-group. The reaction was very positive with 14 organisations willing to participate in the PoC and 16 organisations willing to be part of the Metadata technical sub-group. Organisations were also asked to indicate their willingness to provide in-kind support to the development of the TDO, and despite the fact that the project is still very much in the early stages, 11 organisations indicated a willingness to provide in-kind support (see Appendix 4 for details).

3 Governance structure for the TDO project

The governance structure for the TDO project is illustrated in the graphic below:



The Project Oversight Board is the decision making body for the project. The Advisory Committee is comprised of experts from within the field of data and statistics and in essence their role is to challenge the vision, strategy, implementation plans and objectives for the project, and advise the Project Oversight Board and the Project Team on the key challenges and tasks they identify as relevant to the project. The Project Team, reports to the POB, and is responsible for the day-to-day running of the TDO project.

It is clear that the governance structure, or perhaps membership of the various layers of governance need to evolve as the project progresses. The decision making layer in particular will need to be broadened to reflect the global ambitions of the TDO.

4 Risk Register for the TDO project

A risk register has been developed for the TDO project and is available in excel, and has been provided separately. The register identifies the risks concerned, an assessment of the likelihood of the risk materialising, the impact on the project if the risk materialises and mitigation measures that can be undertaken to minimise or eliminate the impact on the project.

The risks identified were as follows:

- Getting buy-in from key stakeholders (NSOs, IOs)
- Lack of Engagement from Data Compilers
- Standardisation and harmonisation of metadata
- Loss of Leadership / Being Overtaken by Other Initiatives
- Engaging big technology companies to develop their AI algorithms and LLMs to use the TDO
- Current political and geo-political environment
- Creating an active user base
- Sustainability and governance of the TDO
- Inadequate funding to complete the project
- Technological Dependence on External Providers
- Low Usability / Poor Adoption

5 TDO Metadata standard(s) and MVM composition

A group of metadata experts came together on 23 February 2026 to discuss the proposed MVM composition and consider the appropriate metadata standard(s) to be used for the MVM. The organisations represented at this meeting included:

Federal Statistics Office Switzerland, INSTAT Albania, OECD, CSO Ireland, UNDRR, WHO, ISTAT Italy, Liip AG, UNEP-GRID, UNEP, Statistics South Africa, UNCTAD, World Bank, University of Geneva.

The core messages/advice from the group included the following:

- Keep the PoC simple and focus on an Minimum Viable Product
- Develop the TDO concept in stages accepting that there will be challenges and improvements needed over time. Take an agile approach to development
- Limit the scope of the TDO initially to the discovery of indicator data (aggregate statistical data) for the PoC. Linkages to information regarding microdata could be assessed post the initial PoC
- The development of some use cases (set of queries to be put through the various LLMs) was suggested
- The potential of MCP was referenced
- In relation to the metadata standard, an argument was put forward that the standard doesn't need to be limited to just one, as LLMs can read multiple standards. SDMX and DCAT were referenced as suitable standards for use on the TDO
- Optimisation of metadata to support discovery was highlighted as a key initiative. It reflects the importance of using language within our metadata that supports discovery of trusted data. It was emphasised that metadata components need to have semantic value
- The view was expressed that a very small number of metadata components are required strictly for discovery, but these will need to be supplemented by components that fall more into the "filtering components" category (time series, country etc)
- The issue of multilingualism was also raised as being very challenging and for further research post the initial PoC

Reflecting on the input received and setting the feedback in the context of the overall objectives of the project, the composition of the Minimum Viable Metadata set is as follows:

1. Identification and Description

- Title – Clear, descriptive name of the dataset
- Abstract / Description – Short summary explaining what the dataset contains and why it exists
- Keywords / Thematic Tags – To enable domain-based and cross-domain search
- Semantic Topic Tag (from a controlled vocabulary): Essential for accurate semantic search and knowledge graph integration (e.g., using a high-level statistical classification like the UN's Classification of Statistical Activities)
- Unique Identifier – Persistent ID (e.g. DOI, UUID) to avoid duplication

2. Provenance and Trust

- Publisher / Data Provider – Organization or entity responsible for the data
- Date of Publication / Last Update – To assess timeliness
- Source or Origin – Reference to data source (census, survey, administrative, private data holding etc.)

3. Access and Licensing

- Access URL
- Contact Information – For inquiries or additional details.

4. Technical and Structural Information

- Format / Type
- Geographical Coverage
- Temporal Coverage – Time period the data represent
- Units of measurement
- Variable name/label/short description

5. Quality and Usability Signals

- Update Frequency – (e.g., annual, quarterly, monthly, ad hoc).
- Data Quality Note / Methodology Summary – One-line statement of known limitations or quality checks.

6 Proposal for a Proof of Concept

The purpose of the Proof of Concept (PoC) is to validate the feasibility, functionality, and value of the Trusted Data Observatory metadata platform before committing to full-scale development.

It is proposed that a PoC Participating Board be established to oversee/govern the PoC exercise which will be conducted over an agreed period of time decided on by the PoC Participating Board. The composition of the Participating Board should reflect the global ambition of the TDO and include representatives of NSOs and IOs.

The PoC proposal sets out:

- Purpose & Objectives of the PoC
- Scope of the PoC
- Assumptions & Constraints
- Functional Requirements to Validate
- PoC Execution Plan
- Success Criteria & Evaluation Metrics
- Relevant considerations
- Anticipated outcomes

Full details of the PoC can be found in Appendix 5.

The functionality of the TDO, which is to be assessed within the PoC, is set out in the Business Specification for the TDO. Details of the Business Specification can be found in Appendix 6.

7 Institutionalisation, governance, sustainability and hosting

These issues are amongst the most critical issues facing the development and operationalisation of the TDO. The issues of institutionalisation, governance and sustainability are all linked. Solving the institutionalisation problem will contribute to addressing the issues of governance and sustainability.

There have been very many discussions over recent weeks (January/February 2026) with UN organisations on the issues of institutionalisation and hosting, with a number of organisations open, without committing at this point, to the idea of embedding the TDO project within their organisation.

UN organisation can and very often do provide very strong multilateral governance to their activities. It is at the core of how they operate. Governance especially during the development phase of a project often evolves with the project. Given the scope and global ambition of the TDO it will be important to involve representatives from the following sectors in the governance of the TDO at the appropriate stages of the project, and this is generally standard practice for projects led by UN bodies. The future governance composition may benefit from the inclusion of representatives from the following sectors:

- NSOs/IOs
- Researchers
- NGOs
- Policy makers
- Civil Society
- Private sector (need to reflect carefully on this piece)

Developments around institutionalisation and governance are important to address the perspective held by some that the TDO is a Swiss initiative with a Swiss focus rather than a global initiative with global ambitions. At the same time it can also address another perspective which suggests that the TDO is for the Official Statistical Community only. Evolution of the composition of governance boards following institutionalisation and ongoing communication will be important to address these issues.

Institutionalisation can also support sustainability, to a point. Funding may be required to support the development of a Minimum Viable Product and undertaking a Proof of Concept exercise. The level of monetary funding required for these activities may be limited if skilled in-kind contributions based on a multilateral development approach can be taken. However even with in-kind funding it can be anticipated that the hosting organisation and the organisation within which the TDO is institutionalised will require a degree of monetary funding to support their ongoing TDO activities.

In relation to the hosting of the TDO, it is clear that the TDO is to be hosted in Geneva. The hosting environment needs to be cognisant of Data Sovereignty concerns which was an issue raised by very many stakeholders throughout the engagement process.

8 Alignment with other AI related initiatives

Many stakeholders have emphasised the importance of aligning the TDO initiative with other AI related initiatives already underway. In that context the TDO team held discussions with the IMF in New York to explore synergies with the IMF's StatGPT initiative and have agreed to work together throughout March 2026 to explore the opportunities and document the ways in which the two initiatives can complement one another.

The TDO project team have also discussed alignment opportunities with the UNECE HLG MOS project on "AI Ready Dissemination". Written contributions have been submitted to the UNECE to highlight potential synergies and complementarity and a presentation is being given on the TDO to the HLG MOS Blue Skies Network on 19 March, 2026.

The initiatives highlighted above are just two of the AI initiatives underway that provide possible synergies with the TDO project. In an effort to develop greater alignment across global AI initiatives the UNSC at its 57th session in March 2026 decided to establish a city group to address this challenge. The Kigali City Group will operate as an overarching strategic framework to oversee and ensure complementarity in the range of AI Dissemination related initiatives that are underway across the Official Statistical Community. The City Group will aim to strengthen the AI readiness of official data and statistics; ensure that AI systems can reliably discover and properly cite official statistics; enhance metadata standards and promote machine-readable statistical outputs; advance the harmonization and the standardization of metadata and facilitate coordination between ongoing initiatives on AI and statistics. The City group should also provide practical guidance and methodological support to countries. Finally, the group is intended to complement and reinforce existing initiatives, maximising collective impact rather than duplicate them.

9 Conclusions

The Executive Summary sets out the main conclusions.

A lot of progress has been made and stakeholders are very positive about the vision and objectives of the TDO. Of course challenges remain and new challenges will emerge as we move into the next phases of the project which will still involve an important leadership, communications and engagement component, but will also involve an engineering component with the establishment of a physical hosting environment in Geneva (or hopefully alignment with an existing hosting environment) and the development of the platform itself.

Institutionalisation, governance and sustainability are key issues that must be to the forefront of minds as the project moves forward. The key issue is institutionalisation and resolving this issue will go some way to addressing the evolution of the governance of the TDO and also its sustainability.

Engaging the big technology firms, and specifically the developers of the most commonly used LLMs, remains a challenge. Of course the “bottom line” is a concern for them, but protecting and indeed building their brand is also important and supporting the discovery of trusted data sources and engaging in the battle against misinformation and disinformation must be of interest to them.

Appendix 1
TDO Stakeholder mapping exercise

1 Introduction

The TDO is primarily concerned with addressing the issue of Data Discovery in an AI enabled world. It is a platform based on standardised, open, machine readable (AI enabled) metadata to support Discoverability, Comparability, Accessibility and re-use of trusted data sources, for machines as well as persons. The proposed observatory is a metadata platform as distinct from a data platform. There will be no data on the platform which remains within the data compilers' environment and control.

The TDO project is a collaboration between the Data for Change - the PARIS21 Foundation, the Swiss Federal Department of Foreign Affairs, and the Federal Statistical Office (FSO) of Switzerland on a project focussed on enhancing the visibility, discovery and re-use of trusted data.

A critical phase in this project is stakeholder engagement, and this appendix sets out a "stakeholder map" for the project. The key issues covered in this map include:

- Identification of **who the stakeholders are** (individuals, groups, or organizations).
- Understanding **their interests, influence, and impact** on the project.
- Consideration of a **communication and engagement strategy** that fits each stakeholder's needs.

The key Stakeholders

The ultimate audience for the TDO is broad and includes:

- Trusted data compilers
- Governments
- Policy makers
- Non-governmental organisations
- Business Community
- Civil Society – the Public
- The media/data journalists
- Research community
- Academics

The user community also includes those involved in supporting others to find trusted data through lexical and semantic AI-enabled search engines.

However in the context of the TDO project at this point, which is very much at the conceptualisation and early development phase, the key stakeholders include:

- National Statistical Offices (trusted data compilers)
- International Statistical Organisations or International Organisations with a dedicated statistical Division (e.g. World Bank, WHO, IMF, OECD, ILO, UNSD, UNECE). These organisations are both trusted data compilers and data users
- The funders of the TDO project (i.e. philanthropic foundations and the Swiss Federal Department of Foreign Affairs)
- Federal Statistical Office of Switzerland – project initiator and project team members
- Data for Change - the PARIS21 Foundation – project team members
- Geneva Data Community
- The Project Team
- The Project Board

2 Stakeholder Interests, Influence and Impact

The interests, influence and impact of the various stakeholders is summarised in the table below using a combination of project management tools.

The table provides an outline of each stakeholders Power, Influence and RACI status[3] in relation to the project.

The Power–Interest Grid, also known as Mendelow's Matrix, is a stakeholder analysis tool used in project management and change management. Its purpose is to help decide how to engage different stakeholders based on:

- Power: their ability to influence the project
- Interest: how much they care about or are affected by the project

A RACI status is also included for the current phases of the project hat have been funded (phases 1 and 2). Normally a RACI status would be assigned at individual task level and in that context a stakeholder would only ever be one of Responsible, Accountable, Consulted or Informed. However in this context a RACI status has been provided

3 RACI status refers to how clearly roles in a project or process have been assigned using the RACI model. It shows whether each task has defined Responsible, Accountable, Consulted, and Informed roles, and whether any gaps or conflicts remain. A completed RACI status means all roles are agreed upon and aligned among stakeholders

Stakeholder	Involvement	Power	Interest	RACI Status
National Statistical Offices (trusted data compilers)	Trusted data compiler	High	High	Consulted & informed
International Statistical Organisations or International Organisations with a dedicated statistical Division (e.g. World Bank, WHO, IMF, OECD, ILO, UNSD, UNECE	Trusted data compiler and user	High	High	Consulted & informed
The funders of the TDO project	Funder	High	Medium / High	informed
Federal Statistical Office of Switzerland	Project initiator and compiler	High	High	Consulted & informed
PARIS21 Foundation	Data User	Medium	High	Consulted & informed
Geneva Data Community	Data User	Low	High	Consulted & informed
Big Technology Companies	Data distributor	High	Medium	Consulted & informed
Project Team	Project team	High	High	Responsible
Prject Board	Project Board	High	High	Accountable

to give a sense of the role each stakeholder group has in the project and some stakeholders are classified as being both Consulted and Informed.

The table above provides a summary of the relative positioning of each of the stakeholders in the TDO project.

Trusted Data Compilers

Essentially for phases 1 and 2 of the project this cohort includes National Statistical Offices (NSOs) and International Statistical Organisations or International Organisations with a dedicated statistical Division (IOs).

Trusted Data compilers "raison d'etre" is the compilation of high-quality, robust and trusted data to support evidence informed decision making and ensure our citizens have the opportunity to work and live in a society informed by trusted data rather than misinformation and disinformation. It is clear that Data Discovery in an AI enabled world is a key challenge for this stakeholder group.

This is the key group at this stage, as without their involvement the Trusted Data Observatory project can't continue. It is their metadata that will reside on the TDO and provide the roadmap to the relevant trusted data resources.

Ensuring this group understand the rationale for and benefits of the TDO is absolutely essential. Positioning the TDO alongside other developments that are ongoing is also an important task to ensure an understanding that the TDO is supplementing rather than duplicating other activities. In this context embedding the TDO within the suite of other initiatives around Data Discovery and AI readiness of Official Statistics is an important step for the success of the project.

Data Users

The users at this point in the project represent the decision maker group of users. The user community in reality is a very diverse group but the set of users engaging with the project at this stage are those users that recognise the challenges of discovering trusted data especially by LLMs.

Their motivation is primarily altruistic and they recognise the real dangers of misinformation and disinformation.

As we move into the Proof of Concept the user cohort will provide a user perspective of the TDO and help assess the effectiveness of the TDO in addressing their needs.

The Funders

The current Funders of this project include the Swiss Federal Department of Foreign Affairs and two philanthropic foundations. Their motivation is as outlined for the user group above. They are particularly concerned regarding the role of misinformation and disinformation and how it can undermine civic and democratic societies.

This stakeholder group are kept informed on the progress of the project.

Data redistributor – Big Technology

We live in a world saturated with data, where users (and LLMs) do not have the necessary information or indicators to distinguish robust, trusted data, evidence and insight, from poor quality data. In addition the change in user behaviour, in how users find and consume data and information, has evolved with a focus on using AI-enabled search tools where many people want answers quickly, without having to do, or having the time to do, the necessary detailed research (or due diligence) that is often required to identify appropriate sources.

The Big Technology stakeholder group represent the AI community and govern the LLM AI tools that are increasingly being used by users to get answers to their questions. Engaging with this stakeholder group to explore with them how they can support the core objective of the discovery of trusted data and information is an essential part of the project. Understanding their motivations and willingness to support the project is a key task. This is a particularly challenging element of the project but engaging them in a positive way is crucial for the success of the project. Finding common ground and a common interest is at the heart of the task.

Project Board and Team

The day-to-day running of the project rests with the Project Team, which includes a project consultant. The Federal Statistics Office of Switzerland is the key player alongside the consultant and the team member from the PARIS21 Foundation. The Board and Team represent both data compilers and users and so their interests are fully aligned with those set out above for these two stakeholder groups.

3 Resistance and conflict

Broadly speaking the various stakeholder groups are positively disposed towards the concept of the TDO, and acutely aware of the need to tackle the issues of AI readiness and data discovery in that context.

That being said they are a number of areas of concern for some key stakeholders as follows:

Data Compilers:

- Have seen a variety of global initiatives previously that have failed or never been brought to completion and there is always a question regarding projects of this nature within this community
- Governance and sustainability of the TDO are certainly a topic of conversation
- Metadata is complex and the myriad of metadata standards is and always has been a concern, especially for c-suite executives in data compiler organisations
- Positioning within the broader suite of related yet different projects needs careful attention. Focussing on the value-add of the TDO is important
- Data compilers are always concerned about resources and are questioning the level of effort for them in engaging with the TDO

Data Users

- Governance and sustainability into the future is a concern
- Have expressed some element of scepticism regarding the extent to which the creators and owners of LLMs will engage with the TDO and support the overall objective
- Need more time to understand how the TDO works and can support them in their roles

It is perhaps too strong to label these issues as resistance but it probably reflects the type of questioning that one would expect to see in an ambitious, complex and global initiative of this nature. What this points to is the importance of ongoing communication and engagement with these key stakeholder groups.

A website for the TDO has been established and active engagement on LinkedIn is also envisaged to keep stakeholders informed. A number of international events have been arranged on the topic of AI readiness by a range of different bodies and the TDO is an active contributor at all of these events.

A very significant number of bi-lateral and multi-lateral virtual meetings have been held and will continue to be held with the various key stakeholder groups. Both proactive and reactive communication is required for this project and the teams member from the PARIS21 Foundation is leading on this element of the project.

Appendix 2
Stakeholder Questionnaire

Appendix 2 – Stakeholder Questionnaire

Q1) Name of Organization:

Q2) Location/Address:

Q3) Contact Details:

***Name of contact person*:**

***Email address*:**

***Organisation website*:**

Q4) Primary function of your Organisation:

- National Statistical Office
- International Statistical Organisation (including International Organisation with dedicated data/statistical/research team)
- Public Administration including policy development and evaluation
- Non-Governmental Organisation
- Academia/Research
- Media
- Other (please specify):_____

Q5) How familiar are you with the proposal for a Trusted Data Observatory?:

- Very familiar
- Somewhat familiar
- Vaguely familiar
- Not familiar at all

Q6) Based on what you have heard/read etc., what are your primary thoughts/vision in relation to the proposal for a Trusted Data Observatory? Perhaps categorise by 1) What you agree with 2) What you disagree with and 3) What you think is missing

If familiarity is limited please refer to www.trusteddataobservatory.org and circulated communication on the proposal for a TDO

Agree With:

Disagree With:

Issues missing:

Q7) What would you describe as the main benefits of the proposed TDO?

Q8) What risks do you foresee with the TDO project and how can they be mitigated in your opinion?

Q9a) Do you have a dedicated metadata repository currently in place in your organisation?

- **Yes**
- **No**
- **Don't know**
- **Brief description (if Yes)_____**

Q9b) If so, what metadata standards are currently in use?

Q10a) Have you any concerns regarding the proposal to create a Minimum Viable Metadata (MVM) set focussed on metadata elements to support visibility and discovery, and if so can you outline them here?

Q10b) How would you propose that these concerns be addressed?:

Q11) Thinking of the proposal for a Minimum Viable Metadata (MVM) set focussed on metadata elements to support visibility and discovery, what elements do you see as being essential for inclusion on the MVM, keeping in mind the intention to create a minimal set?

Q12) Can you set out your thoughts on the potential use cases identified for the TDO which include:

Primary use cases

Supporting Large Language Models (LLMs) with Verified Data - to enable LLMs to answer factual questions using verified, up-to-date data sources

Mapping Existing Data Sources - To provide an overview of what data exists, in which domains, and from which sources.

Monitoring Data Availability and Quality - To indicate whether data sources remain accessible and whether their quality is maintained over time

Secondary use cases

International Collaboration - Multilateral organizations can use TDO to coordinate data efforts and harmonize standards

Crisis Response - Rapid identification of available data during emergencies

Data Discovery for Policy Design - Policymakers can identify gaps in data availability across regions or sectors

Educational Use - Teachers and students can explore real-world data sources for learning and projects

1) What use cases do you agree with 2) What use cases do you disagree with and 3) What use cases do you think might missing

Agree With:

Disagree With:

Use cases missing:

Q13) Would you or your organisation be willing to participate in the technical group that will be established to identify the MVM and associated harmonised metadata standard?

- **Yes**
- **No**
- **Don't know yet**

Q14) The feasibility report from November 2024 outlined some initial thoughts on the functionality of the TDO, including:

- ***API enabled extraction and receipt of harmonised and structured metadata***
- ***Provision of metadata that is machine readable and AI enabled supporting lexical and semantic search functionality***
- ***Once on platform the ability to refine searches by pre-defined filters (e.g. theme, data compiler type)***
- ***The ability to monitor usage of the platform***
- ***Feedback mechanism to support ongoing development and refinement***
- ***User friendly interface***
- ***Usage of AI tools on the platform to prepare reports on comparability of data holdings for a specified search (advanced – downstream iterations)***

Reflecting on the above list, can you outline 1) What you agree with 2) What you disagree with and 3) What functionality you think is missing?

Agree With:

Disagree With:

Issues missing:

Q15) Phase 3 of the TDO project envisages a Proof of Concept (PoC) exercise. Would your organisation be willing to participate in the PoC? *(Note: For data compilers/producers this would involve implementation of the MVM and associated standard and trialling the functionality of the TDO, including the automatic loading and extraction of the MVM)*

- **Yes**
- **No**
- **Don't know yet**

Q16) The long-term governance of the TDO needs consideration. It will be a global metadata platform and in that context long-term governance structures will need to reflect that global dimension. Have you any thoughts on the future governance of the TDO and how it should be structured and supported?

Q17a) Are you willing to provide in-kind support during the development and implementation phases of the TDO project?

- **Yes**
- **No**
- **Don't know yet**

Q17b) If Yes, in what key areas do you think you can provide support. Please specify:

Q18) The long-term sustainability of the TDO needs consideration. Have you any thoughts on how the TDO can be sustained post the development and initial implementation of the TDO?

Q19) The focus of the project is on the discovery, visibility and by extension re-use of TRUSTED data sources. In the initial phases of the project the TDO will focus on the data holdings of National Statistical Offices (NSO's) and International Statistical Organisation or International Organisations with dedicated statistical divisions (IO's). Looking beyond NSI's and IO's, have you any thoughts on the criteria that could be used to identify other trusted data sources?

Q20) Looking at the metadata expertise available to you, do you see a capacity building requirement for your organisation moving forward?

Q21) Are there any additional elements that would encourage your or you organisation to engage with and participate in the TDO? Please feel free to include any comments or thoughts you may have.

Q22) Are you aware of any other initiatives similar to the TDO proposal, or any other initiatives that he TDO should be aligned with

Yes (please specify):_____

No

Don't know

Many thanks for taking the time and effort to provide your input on the Trusted Data Observatory. All inputs will be reviewed and will inform the various phases of the TDO project.

Appendix 3 Frequently Asked Questions (FAQs) Document

What is the rationale behind the Trusted Data Observatory?

The TDO is primarily concerned with addressing the issue of Data Discovery in an AI enabled world. It is a platform based on standardised, open, machine readable (AI enabled) metadata to support Discoverability, Comparability, Accessibility and re-use of trusted data sources, for machines as well as persons. The proposed observatory is a metadata platform as distinct from a data platform. There will be no data on the platform which remains within the data compilers' environment and control.

Is the key to Data Discovery not just about optimising our websites and improving our Data Platforms, ensuring our data is Open?

In short, no. Despite the fact that trusted data compilers e.g. (NSOs and International Organisations) already compile high-quality trusted data, in many cases adhering to Linked Open Data standards, with associated metadata, on optimised web sites, simple searches using standard search tools, or indeed AI supported search tools (ChatGPT, Gemini, Copilot etc), often provide plausible, yet inaccurate results. These LLMs are not discovering our data even with all of this in place.

So if the answer isn't the provision of better data platforms, optimised websites and the provision of Open Data, what is?

Metadata is the key. LLMs are not good at engaging with numbers, but they can engage with metadata.

If metadata is the answer, why are trusted data sources that already have high-quality well-structured metadata associated with those sources in place not being found/discovered as often as one would expect?

Even where well-structured high-quality and comprehensive metadata is in place, it is not regularly open, visible or machine readable, and the international metadata standards used varies (e.g. Dublin Core, DDI, SDMX, DCAT).

How does our approach to metadata need to change?

Traditionally the focus of trusted data compiler's has been on the data, with the organising and structuring of the metadata seen as secondary in some way. Access to the metadata, has pretty much been facilitated by first of all accessing the data. Our paradigm has very much been Data to Metadata. In an AI enabled world, where our first task is to ensure our data is discovered (if it's not discovered it certainly won't be understood), that paradigm needs to be inverted to Metadata to Data.

What additional value does the TDO provide if we all just focus on improving our metadata nationally within our own environments?

Even if we could all agree on one metadata standard that is designed to support Data Discovery in an AI enabled world, experience tells us that decentralised implementation of internationally agreed standards often leads to slight or sometimes significant differences in implementation at national level. National specificities, cultural and perhaps institutional settings/environments can impact on implementation and in that context differences arise.

What is the Minimum Viable Metadata (MVM) set?

The Minimum Viable Metadata set, or indeed it could be referred to as the "Minimum Discovery Metadata set", **is a subset** of the detailed metadata holdings of trusted data compilers. The MVM refers to those metadata elements that support data discovery. It does not involve the creation of new metadata elements or components.

Does the introduction of the MVM represent the minimum suite of metadata we need as trusted data compilers?

No. As outlined above the MVM is a subset of the detailed metadata holdings of trusted data compilers. The existence of high-quality, well-structured, detailed metadata is a pre-requisite for all trusted data compilers.

Will the introduction of the MVM mean the development of yet another metadata standard?

No. The standard to be used will be one of the existing metadata standards that are considered to be AI enabled.

Does the focus on the MVM set not represent a loss of information on key issues required to assess quality, methodology, support interpretation and interoperability?

The objective of the TDO is primarily to support data discovery. The TDO only requires those metadata elements that support discovery, and components focussed on

methodology for example do not support discovery, indeed it could be argued that it may complicate discovery. Once a data sources has been "discovered" through the TDO, the user will then be able to access the detailed metadata, including methodology focussed elements, at source via the link provided through the TDO.

Will the TDO address underlying data problems like definitional differences or use of different classifications for variables?

It is not within the scope of the TDO project to address underlying data problems such as inconsistency in code lists used for certain variables like sex and marital status, definitional differences, data quality etc.. Of course these types of issues are problematic and certainly need to be addressed but those issues need to be addressed elsewhere. If we attempt to use the TDO to resolve not just data discovery issues but the broader suite of underlying data issues, then the TDO project will not succeed.

We have limited resources, so how much work will be involved for NSOs to engage with the TDO?

For those trusted data compilers that already have high-quality, well-structured, and detailed metadata, the effort will be minimal. The MVM is a subset of the existing metadata, so no additional work is required to develop the MVM, once the detailed metadata is in place. It is intended to automate the process of re-classifying the metadata standard to the selected AI enabled standard, should it be different, in the process of loading the metadata to the TDO.

What does the "journey" to the TDO look like for a National Statistical Office?

As always it depends on your starting point. As outlined above where high-quality, well-structured, and detailed metadata is already in place and available through a metadata catalogue then the work load will be minimal.

Who decides what trusted data is?

This is a key question and relates to the ultimate governance of the TDO. The observatory is designed to be 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 but there is a clear recognition that ultimately the governance of this global observatory, even though it will be located in a global organisation based in Geneva, must be governed at the global level. It may be useful to reflect on the role the community of Chief Statisticians of NSOs 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).

What is the definition of trusted data?

Initially, at least, the TDO will limit its scope to the data holdings of National Statistical Offices (NSOs) and the holdings of International Statistical Organisations or International Organisations with recognised statistical Divisions (IOs). Such organisations compile their statistical data within the framework of internationally agreed principles (e.g. UN FPOS, ES CoP), statistical standards and classifications.

It is important to acknowledge that there are trusted data sources beyond NSOs and IOs, and the inclusion of these sources is certainly something that will need to be explored as the TDO develops and evolves over time. Criteria will be developed in collaboration with stakeholders to identify trusted data compliers and their sources.

What about Data Comparability? How can and will the TDO address this?

Not directly, but indirectly. The focus of the TDO is supporting the discovery of existing trusted data sources "as they are". Data quality and comparability are perennial challenges and have and will continue to be a focus for trusted data compilers for many years to come. The TDO, by supporting discovery, will indirectly highlight quality and comparability issues and therefore in that context play a role in helping to highlight where these issues arise.

How will the TDO address Data Security and Data Privacy concerns

Three considerations:

1) It is important to remember the TDO is a metadata platform and not a data platform. Metadata can on occasion incorporate sensitive information which requires a privacy and security lens. However the metadata to be included on the TDO focusses only on metadata designed to support discovery which in general will not contain any sensitive information

2) The focus of the TDO is supporting the discovery of trusted data that is already in the public domain – published in statistical reports, releases, press conferences, infographics, visualisations, aggregate data cubes etc.

3) Finally the TDO will not provide access to microdata. It is intended that the TDO will make users aware of the existence of microdata files and link to information that outlines to users the criteria and requirements for access to these microdata files, including if access is possible

How will the TDO engage with existing national metadata catalogues?

A push and pull approach will be used for loading the MVM set to the TDO. APIs will automatically extract metadata from existing national metadata catalogues and automatically load this metadata to the TDO. It is intended that the functionality of the TDO will also include a metadata editor that will take metadata in whatever internationally agreed metadata standard is in use in the national metadata catalogue and convert it to the agreed MVM AI enabled metadata standard. Trusted data compilers will also be provided with the option to load their metadata to the TDO directly themselves.

Finally the code underpinning the TDO will be open source in nature and this will be available to all users, including trusted data compilers to replicate at a national level should they wish to do so.

What does implementation of the TDO look like from a technical perspective?

At a high level, one could envisage an initial Trusted Data Observatory being built as a lightweight metadata discovery platform. For a Minimum Viable Product (MVP) or proof of concept, this could be done by partnering with a small number of trusted data producers at least one National Statistical Office and one International Organisation and pulling in links and core metadata from their existing catalogues. The initial technical scope could be kept tight by defining a simple, standards-aligned metadata profile (for example, a DCAT) that covers the key fields such as dataset description, producer, update frequency etc. This metadata could then be ingested through straightforward API calls, normalised, and stored in a central metadata database to allow indexing and basic validation. A simple search interface and API would make this catalogue accessible, with the Observatory acting primarily as a trusted pointer to authoritative data sources rather than trying to host the data itself.

Looking beyond the MVP, there is a clear opportunity to introduce semantic search, which really gets to the heart of what the Trusted Data Observatory is trying to solve helping people quickly find the right trusted data across institutions and borders. By building a semantic index over validated metadata and grounding everything strictly in what is actually provided by official catalogues and access endpoints the Observatory could support intent-based searching in a way that dramatically reduces the time users spend trying to locate the information they need. In addition, we could also bring in the Model Context Protocol (MCP) so that the Observatory becomes directly discoverable by AI systems and agents. This would allow AI tools to query and retrieve official metadata through a controlled, authoritative interface, improving machine-to-machine findability while keeping the Observatory firmly anchored as a standards-based, trusted discovery layer.

Appendix 4
TDO Participation Lists

MVM technical sub-group

UN WHO, Vietnam, Albania, Rwanda, UNEP, Ireland, Finland, OECD, South Africa, Italy, UNCTAD, Switzerland, World Bank, UNDRR, Liip, Portugal

Proof of Concept participants

Albania, WHO, Vietnam, UNEP, Ireland, OECD, Lithuania, UNCTAD, World Bank, Rwanda, Switzerland, Liip, UNDRR (likely), Portugal

Provision of in-kind support

Vietnam, Albania, UNEP, OECD, South Africa, WHO, UNCTAD, World Bank, Rwanda, Switzerland, Liip AG

Appendix 5
Proof of Concept proposal for the TDO

1 Purpose & Objectives of the PoC

1.1 Purpose

The purpose of this Proof of Concept (PoC) is to validate the feasibility, functionality, and value of the Trusted Data Observatory metadata platform before committing to full-scale development.

1.2 Objectives

The PoC aims to:

- Validate the role of the TDO in supporting the discovery of trusted data sources in an AI enabled world
- Focus on the discovery of aggregate statistical data (statistical indicators) that are already in the public domain
- Validate the use of the Minimum Viable Metadata set, a sub-set of the detailed metadata holdings of trusted data compilers
- Assess the automated transformation of internationally agreed non-AI enabled metadata standards to an AI enabled metadata standard used for metadata on the TDO
- Assess the challenges and resources involved for trusted data compilers to engage with the TDO
- Assess the use of API's to support interaction with the TDO
- Evaluate scalability, extensibility, and performance
- Identify risks, gaps, and implementation considerations

2 Scope of the PoC

2.1 In Scope

The PoC will cover:

- The Discovery of publicly available trusted data (aggregate data – indicators) using LLMs (ChatGPT, Copilot, Gemini etc.)
- The Discovery of publicly available trusted data (aggregate data including indicators) by humans on the TDO platform itself
- Automated extraction of metadata from national/IO compilers environment to the TDO platform
- Automated "pushing/loading" of metadata from the trusted data compilers environment to the TDO
- A limited set of data domains e.g. Inflation, economic indicators, will be tested in the PoC
- A defined subset of metadata components (MVM)
- The suitability/functioning of the hosting environment
- The functionality of the TDO
- The effectiveness of the selected AI enabled metadata standard(s)

2.2 Out of Scope

The PoC will **not** include:

- Addressing underlying data issues (e.g. definitional differences, classification issues) as this is outside of the intended scope of the TDO itself
- Trusted data holdings of Non-NSI/IOs
- Assessment of criteria to be used to determine what constitutes "trusted data"
- Geo-spatial data, other than that already incorporated within statistical data holdings
- Supporting the development of national metadata catalogues to feed the TDO
- User behaviour in terms of mechanisms used to discover trusted data

3 Assumptions & Constraints

3.1 Assumptions

- Good quality metadata (MVM), classified to a recognised international standard is available in a metadata catalogue held by the trusted data compiler
- The trusted data compilers technology environment supports the use of APIs to support machine-to-machine exchange of metadata
- The availability of a fully functioning metadata editor that can transform one internationally agreed metadata standard to another internationally agreed metadata standard (e.g. DDI to SDMX or SDMX to DCAT)
- Constructive engagement from then Open AI community i.e. developers of LLMs

- Required access to selected systems is available
- The hosting environment has the necessary technology environment in place for the PoC including permissions and access rights
- Stakeholders are available to participate, review and feedback
- Constructive engagement with Open AI community, LLM developers, to ensure usage of the TDO by LLMs to support discovery of trusted data sources has taken place

3.2 Constraints

- The work of the HLG MOS group on "AI-Readiness: Dissemination" may not be available in time for the PoC, and therefore the learnings from their work may not be reflected in the version of the TDO platform used for the PoC
- The "institutionalisation" of the TDO may not be completed in advance of the PoC, and therefore the engagement processes and procedures may be subject to change post PoC to reflect the new environment
- The timing and duration of the PoC may be subject to the availability of the organisations participating. Ideally all testing should take place at the same time. Every effort will be made to ensure this happens
- The hosting environment (infrastructure) used in the PoC may reflect more of a "sandbox" environment rather than a "live" environment, depending on the hosting organisation
- Simplified governance and approval processes will be in place for the PoC

4 Functional Requirements to Validate

- Automated harvesting of metadata from selected sources (e.g., NSO metadata catalogues) within acceptable time limits
- Automated "pushing/up-load" of metadata from selected sources to the TDO platform, within acceptable time limits
- Manual metadata entry or enrichment
- An AI interface based on the MCP protocol (may not be in place in time for the PoC)
- Automated transformation, using a metadata editor, of metadata to the agreed metadata standard for the TDO where the ingested metadata is in a different international agreed standard
- Advanced search functionality to support data discoverability:
  - 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)
  - 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) should be included
  - Facilitation of LLM queries

  Note in all three cases above the resulting response must provide a direct, resolvable link back to the trusted data compiler source (and by extension the underlying detailed metadata in the compiler's environment)

- Compliance with the Minimum Viable Metadata (MVM) set, which is essentially metadata components that supports data lineage and data discovery. These components will largely be available from trusted data compilers reference metadata. A lightweight mechanism for institutions to validate or "approve" the metadata they have contributed
- "Filtered" search functionality based on specified categories including for example, data sources (e.g. NSI data only), themes (e.g. SDG's, agriculture, employment), frequency, periodicity etc.
- Assessing a "feedback" option to users to support the ongoing development and evolution of the observatory
- Usage monitoring and the provision of metrics to improve functionality over time
- Metadata versioning and update history – "last updated" indicator
- Metadata browsing and navigation
- Intuitive user interface for human users of the TDO that is responsive
- Ability to scale to additional data domains
- Integration with future tools (e.g. chat bots on trusted data compiler's environments) should be considered in TDO design
- Metadata access controls (high-level) for administrators and metadata contributors are in place

In relation to the PoC, the functionality that can be tested/assessed will be limited to the functionality built into the beta TDO version. As the TDO functionality evolves the new elements will need to be appropriately tested.

5 PoC Execution Plan

- A PoC Participating Board will be established to oversee/govern the PoC exercise
- The PoC will be conducted over an agreed period of time decided on by the PoC Participating Board
- Experts from the participating organisations will agree a suite of data queries for the selected domains
- Each organisation participating in the PoC will submit these queries over an agreed time period across an agreed number of consecutive days, prior to launch of the TDO to assess LLM performance in the absence of the TDO
- The results of these searches will be captured by each participating organisation including the sources used in the provision of the response from the LLMs and any variation in responses from the LLMs across the "testing" days
- The results will be transmitted to the PoC coordinator and PoC Participating Board
- Metadata from the participating organisations will be ingested/uploaded to the TDO environment
- Metadata experts will review the ingested metadata to assess quality. Compliance with the MVM and agreed AI enabled metadata standard should be ensured through the automated processes
- Participating organisations will repeat the data queries outlined above and assess the impact of the TDO on data discovery
- The results of these searches will be captured by each participating organisation including the sources used in the provision of the response from the LLMs and any variation in responses from the LLMs across the "testing" days
- The results will be transmitted to the PoC coordinator and PoC Participating Board
- Manual users will conduct the same searches on the TDO through the search tools provided on the TDO
- The filtering functionality on the TDO should also be tested using queries developed by the PoC Participating Board

6 Timeline for the PoC

The exact timing of the PoC will be determined by the Project Oversight Board in consultation with the project team and PoC participating organisations via the PoC Participating Board.

7 Success Criteria & Evaluation Metrics

7.1 Data Discovery

This is the key test from the PoC. The TDO will be assessed in terms of its ability to discover trusted data sources.

Evaluation Metric: For each of the data queries conducted prior to the launch or operationalisation of the TDO, the percentage of trusted data sources referenced in responses or search results for LLM, lexical and semantic searches will be calculated.

The same percentages will be calculated post launch or operationalisation of the TDO, and the percentages will be compared across each search by organisation.

7.2 Functionality testing

The testing will include an assessment of each of the proposed functions of the TDO as set out in section 4 above e.g. metadata successfully ingested from all PoC sources, compliance with the MVM requirements, feedback mechanism operating as expected, metadata transformation successful.

Success Metric: The evaluation metric will be straight forward in that the function either did or did not successfully perform the function.

7.3 Architecture/Tech environment test

The architecture and technology environment will also be assessed to ensure it provided a responsive service to users.

Evaluation Metric: Response times to queries submitted are within normal standards expected. No downtime in the host environment experienced. No access limitation reported by users.

7.4 User Friendly Interface

The User Interface for human interaction on the TDO will also be reviewed.

Evaluation Metric: Human users interacting with the TDO (either searching for trusted data sources or directly uploading or updating previously loaded metadata) find the TDO easy to navigate and responsive. A qualitative report will be provided by users outlining areas or satisfaction and areas for further development.

8 Considerations

The scope or functionality of the TDO for the PoC has yet to be finalised. The UNECE HLG MOS project on AI-Ready Dissemination may not have fully reported or completed their work in advance of the PoC and therefore the learnings from that project may not be factored into the TDO for the PoC.

Engagement with the Open AI community and the community of LLM developers is proving challenging. The PoC will need to await the outcome of engagements with these developers to fully understand and support the design of the TDO to ensure it is AI enabled and machine readable.

As outlined elsewhere the institutionalisation of the TDO is an important next step. The Project Oversight Board need to consider if the PoC can proceed in advance of this issue being resolved.

A decision on the hosting environment for the TDO also needs to be made. That said, a number of organisations have already indicated their willingness to host the TDO for the PoC.

9 Anticipated outcomes

A full report of the PoC should be prepared focussing on the issues outlined in section 7 - Success Criteria & Evaluation Metrics.

In addition, the findings from the PoC should be captured and used as a key input to refine further development of the TDO including a “go/no go” decision on further development.

Appendix 6
TDO Business Specification

What is the Trusted Data Observatory?

The TDO is primarily concerned with addressing the issue of Data Discovery in an AI enabled world. It is a platform based on standardised, open, machine readable and AI enabled metadata to support Discoverability and Transparency. The focus of the TDO is supporting the discovery of existing trusted data sources "as they are". The TDO, by supporting discovery, will indirectly highlight quality and comparability issues and therefore in that context play a role in helping to highlight data issues.

The proposed observatory is a metadata platform (i.e. an indexed metadata catalogue) as distinct from a data platform, that incorporates access endpoints that link directly to the relevant data source. There will be no data on the platform which remains within the data compilers' environment and control.

The TDO will provide links to the following through the metadata on the TDO:

- Trusted data (aggregate statistical indicators) that is already in the public domain – published in statistical reports, releases, press conferences, infographics, visualisations, aggregate data cubes etc.
- The existence of microdata holdings – the TDO will not provide access to microdata. It is intended that the TDO will make users aware of the existence of microdata files and link to information that outlines to users the criteria and requirements for access to these microdata files, if access is possible. An example of what the TDO should link to in relation to microdata can be found here: Data for Researchers – CSO – Central Statistics Office

Metadata

Most of the internationally agreed metadata standards in use (e.g. Dublin Core, DDI) were developed before the evolution of AI, and in that context many of the standards in use are generally not AI enabled. The function of metadata traditionally has been to help users to understand and interpret data but not discovery in an AI enabled world. The AI enabled metadata[1] on the TDO will be classified according to the DCAT standard and SDMX. The specific standard in itself is not the important issue. What is important is that the metadata is available in a structured and consistent machine readable format that can be interpreted and understood by LLMs.

The concept of a Minimum Viable Metadata (MVM) set, or indeed it could be referred to as the "Discovery Metadata set", is being introduced with the TDO. The MVM is a subset of the detailed metadata holdings of trusted data compilers. The MVM refers to those metadata elements that support data discovery. It does not involve the creation of new metadata elements or components.

By focussing on the MVM we can streamline the number of elements that need to be ingested and reviewed, and in that context simplify and support discovery specifically.

Trusted Data

Initially, at least, the TDO will limit its scope to the data holdings of National Statistical Offices (NSOs) and the holdings of International Statistical Organisations or International Organisations with recognised statistical Divisions (IOs). Such organisations compile their statistical data within the framework of the UN Fundamental Principles of Official Statistics and in line with internationally agreed statistical standards and classifications.

How will the TDO engage with existing national metadata catalogues

A push and pull approach will be used for loading/extracting the MVM set to the TDO. API's will automatically extract ("pull") metadata from existing national metadata catalogues and automatically load this metadata to the TDO. It is intended that the functionality of the TDO will also include a metadata editor that will take metadata in whatever internationally agreed metadata standard is in use in the national metadata catalogue (e.g. Dublin Core, DDI) and convert it to the agreed MVM AI enabled metadata standard(s) on the TDO).

[1] In practice, an AI-enabled metadata standard means the metadata isn't just a static checklist of fields – it's designed so AI can create, interpret, enrich, and act on it reliably

Trusted data compilers will also be provided with the option to load their metadata to the TDO directly themselves – a "push" facility.

Finally the source code underpinning the TDO should be open source in nature and this will be available to all users, including trusted data compilers to replicate at a national level should they wish to do so.

A high-level technical perspective of the platform

At a high level, one could envisage an initial Trusted Data Observatory being built as a lightweight metadata discovery platform. For a Minimum Viable Product (MVP) the initial technical scope could be kept tight by defining a simple, standards-aligned metadata profile (for example, a DCAT) that covers the key fields such as dataset description, producer, update frequency etc. This metadata could then be ingested through straightforward API calls, normalised, and stored in a central metadata database to allow indexing and basic validation. A simple search interface and API would make this catalogue accessible, with the Observatory acting primarily as a trusted pointer to authoritative data sources rather than hosting the data itself.

However we need to look beyond a MVP. There is a clear requirement to introduce a semantic search facility, which gets to the heart of what the Trusted Data Observatory is trying to solve – helping people quickly find the right trusted data across institutions and borders. By building a semantic index (layer) over validated metadata and grounding everything strictly in what is actually provided by official catalogues and access endpoints, the Observatory would support intent-based searching in a way that dramatically reduces the time users spend trying to locate the information they need. In addition, the Model Context Protocol (MCP) should be incorporated so that the Observatory becomes directly discoverable by AI systems and agents. This would allow AI tools to query and retrieve official metadata through a controlled, authoritative interface, improving machine-to-machine findability while keeping the Observatory firmly anchored as a standards-based, trusted discovery layer.

Hosting environment

The infrastructure required is relatively lightweight as the amount/extent of processing will be limited. The key question to consider is whether to run the platform on Windows or Linux. The latter is definitely more cost efficient and when starting from scratch it's more of an option. Depending on hosting providers it may be possible to use cheaper non-MS related services for the Database and storage. For the metadata providers the opportunity may exist to utilise their existing web presence to either host the data or provide the gateway.

TDO Functionality

- Automated harvesting of metadata from selected sources (e.g., NSI/IO metadata catalogues) within acceptable time limits
- Automated "pushing/up-load" of metadata from selected sources to the TDO platform, within acceptable time limits
- Manual metadata entry or enrichment
- An AI interface based on the MCP protocol
- Automated transformation, using a metadata editor, of metadata to the agreed metadata standard(s) for the TDO where the ingested metadata is in a different internationally agreed standard
- Advanced search functionality to support data discoverability:
  - 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)
  - 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) should be included
  - Facilitation of LLM queries

  Note in all three cases above the resulting response must provide a direct, resolvable link back to the trusted data compiler source (and by extension the underlying detailed metadata in the compiler's environment)

- Compliance with the Minimum Viable Metadata (MVM) set, which is essentially a set of metadata components that supports data lineage and data discovery. These components will largely be available from trusted data compilers reference metadata. A lightweight mechanism for institutions to validate or "approve" the metadata they have contributed
- "Filtered" search functionality based on specified categories including for example, data sources (e.g. NSI data only), themes (e.g. SDG's, agriculture, employment), frequency, periodicity etc.
- Assessing a "feedback" option to users to support the ongoing development and evolution of the observatory
- There is strong demand for a "multilingual" TDO. It is proposed to develop an English version for the PoC, and depending on the outcome, the multilingual requirement can be further assessed
- Usage monitoring and the provision of metrics to improve functionality over time
- Metadata versioning and update history – "last updated" indicator

- Metadata browsing and navigation
- Intuitive user interface for human users of the TDO that is responsive
- Ability to scale to additional data domains
- Integration with future tools (e.g. chat bots on trusted data compiler's environments) should be considered in TDO design
- Metadata access controls (high-level) for administrators and metadata contributors
- Built to support potential inclusion of a "chatbot" similar to the IMF's "StatGPT" bot

Technical specification for developers

This business specification needs to be used to prepare a detailed technical specification for developers, system architects, and technical project managers. This will require Business Analysts and software engineers and architects.