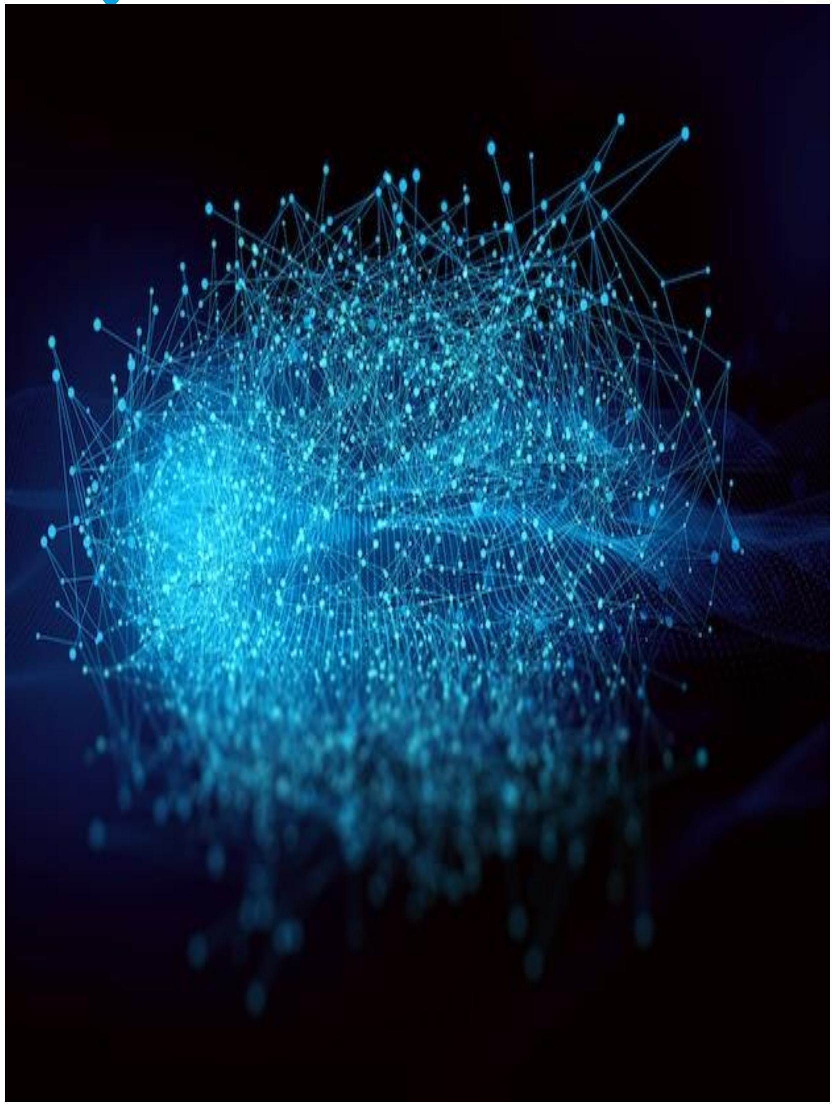

ITU Workshop “At the crossroads of Standards and Research: AI/ML datasets for future networks”, Geneva, 16 July 2024

Overview of the ITU-T Correspondence Group on datasets applicable for AI/ML in networks (CG-datasets)

Marco Carugi and Vishnu OV Ram
ITU-T CG-datasets co-convener
marco.carugi@gmail.com , vishnu.n@ieee.org

AI/ML at ITU-T SG13

- Artificial Intelligence (AI)/Machine Learning (ML) studies are now pervasive across all ITU-T SGs
- ITU-T SG13 deals with Future Networks and addresses in its mandate the application of AI/ML technologies in networks (requirements, capabilities, architectural frameworks, functions)
- A lot of of AI/ML related specifications have been published by SG13 since the foundational Rec. Y.3172 ("*Architectural framework for machine learning in future networks including IMT-2020*") approved 03/2019
- Focus Groups activities [FG-ML5G (2018-11 to 2020-07) and FG-Autonomous Networks (2020-12 to 2024-03)] have also contributed - significantly - to the development of AI/ML related specifications in SG13
- And SG13 is actually considering the establishment at this July 2024 SG13 meeting of a new Focus Group related to AI: FG-AINN (*Artificial Intelligence Native for Telecommunication Networks*)
- AI/ML is acquiring a central role in Future Networks (Beyond IMT-2020, IMT-2030) and consequently in the network related standardization activities such as those of ITU-T SG13
- In this dynamic context, SG13 established in 2022 the ITU-T Correspondence Group on "Datasets applicable for AI/ML in networks"

Why the topic of datasets applicable for AI/ML in networks is important

- The use of AI/ML technologies is having a considerable impact on networks, with data-driven ML methods enabling fundamentally new intelligent design and decision-making in networks.
- However, the advancements with ML for networks relies on the availability of datasets to test the results and attempts generalizations. These datasets serve as the foundational bedrock upon which algorithms are trained, validated, and tested, playing a critical role in the development and advancement of AI/ML models. Their importance is multifaceted, encompassing aspects of data quality, model performance, reproducibility, and innovation in the field. Established datasets should reflect real-world scenarios and problems, providing a practical framework for developing and testing AI/ML solutions.
- One of the main bottlenecks is the current limited availability of datasets from either practical simulations or experimental testbeds that can be considered as reference datasets. Creating reference datasets is expensive, while the commercial datasets from telecommunication operators are mostly inaccessible.
- It is then important to progress the studies on standardization requirements of datasets applicable for AI/ML in networks and identify directions for standardization approach for these datasets.

ITU-T Correspondence Group on datasets applicable for AI/ML in networks (CG-datasets)

Established by ITU-T SG13, it began its activities in Nov 2022, its life time expires at the July 2024 SG13 meeting

- CG-datasets: **https://www.itu.int/en/ITU-T/studygroups/2022-2024/13/Pages/Correspondence-Groups.aspx**

Co-conveners: Vishnu Ram OV (Independent Expert - vishnu.n@ieee.org), Marco Carugi (Consultant, Huawei Technologies - marco.carugi@gmail.com)

Open to ITU members and non-ITU members (non-ITU members' access is moderated by the co-conveners)

CG mandate in summary (Terms of Reference contained in SG13-TD151/PLEN)

- To study standardization requirements of datasets applicable for AI/ML in networks
- To conduct a gap analysis of related standardization efforts
- To identify directions for standardization approach for datasets applicable for AI/ML in networks
- To develop a Technical Report with technical insights and recommendations to ITU-T SG13 for a standardization approach for datasets applicable for AI/ML in networks

Communication and collaboration channel with SDOs and industry bodies

- No formal liaisons, but participation of ETSI ISG ENI Chair and ATIS ANA co-Chair in numerous CG meetings

CG-datasets working methods

- **CG mailing list: cgdatasets@lists.itu.int** (>50 subscribers)
- **Mailing list subscription:** "Login" button at **https://www.itu.int/en/ITU-T/ewm/Pages/services.aspx**

  [Prerequisite: ITU User (TIES/Guest) account: https://www.itu.int/en/ties-services/Pages/login.aspx]
- **Sharepoint repository for the CG documents**: **https://extranet.itu.int/sites/itu-t/studygroups/2022-2024/sg13/cg-datasets/SitePages/Home.aspx**
- **Online meetings (2 hours approx. every month)**
  - 16 official CG meetings (and 6 restricted meetings), the last CG meeting was on 26 June 2024
  - review of input contributions, progress of the deliverable, action points and future meetings
- **CG progress reports provided to ITU-T SG13 [**CG reports provided to the SG13 meetings held in March 2023, Oct 2023, March 2024 and July 2024]

CG-datasets deliverable

Technical Report (TR) on “Datasets standardization approaches for datasets applicable for AI/ML in networks – First Edition”

- **Information collected from market and research experiences on datasets, models and tools (as applicable for 5G and future networks) concerning different use case domains**
- **Use case domains considered in this First Edition**:
  - Data-driven Network Optimization
  - Telecom network operations supported by Large Language Models
  - Autonomous Networks (AN) - in collaboration with ITU-T Focus Group on AN (FG-AN) and with consideration of Suppl. 71 to ITU-T Y.3000 series [“Use cases for autonomous networks”]
  - End-customer engagement and causal analysis
  - Reconfigurable Intelligent Surface (RIS)-aided Vehicle-to-Everything (V2X)
- **According to the use cases (domains), the Report elaborates on relevant datasets, models, toolsets, analysis and remarks, recommendations for future standardization and research activities**
- **The final version of the TR - 13 July 2024 - is accessible from the CG-datasets page (see also the link in the last slide of this presentation)**

Communication and collaboration of the CG-datasets with other expert groups

ITU-T FG-AN

- Three joint sessions held between CG-datasets and FG-AN in the context of three FG-AN meetings (20 April, 13 July and 28 Sept 2023), with the aim to discuss matters of common interest
- FG-AN has contributed a set of CG-relevant AN use cases to CG-datasets (with datasets, models and other associated information) - as well as the CG-datasets has contributed other AN use cases to FG-AN

ETSI ISG ENI (Industry Specification Group Experiential Networked Intelligence)

- ENI activities overview, ENI System Architecture and Models (presented by ENI Chair and Architecture Rapporteur)

ATIS ANA (AI Network Applications)

- ATIS ANA co-convener has participated (and contributed) in various CG calls, CG-datasets leaders joined the 4 Dec 2023 ANA e-meeting for information exchange and possible future collaboration

ITU-T SG13 experts

- Various SG13 experts have participated in the CG-datasets activities

Final report of the CG-datasets to the 15-26 July SG13 meeting: the actions for the parent SG13

1. To note the final report and the results achieved by the CG-datasets.
2. To note the results of today’s ITU-T workshop
3. To approve the Technical Report on “Datasets standardization approaches for datasets applicable for AI/ML in networks – First edition”
4. To consider, and hopefully approve, the request for extension for one year of the life time of the CG-datasets (updated ToR provided in the Final Report)
5. To consider the CG-datasets suggestions for ITU-T SG13, including possible standardization study proposals that might be identified by ITU-T SG13 in relation to the findings of the Technical Report – First Edition (in particular those indicated in clause 7). The participants of the CG-datasets activities are open to support possible standardization studies.

Structure of the CG Technical Report

- **Introduction (clause 1) -** background, existing work in the area of datasets and motivations for studying reference datasets
- **Use cases (clause 2) –** relevant telecom domain use cases which use datasets
- **Datasets for AI/ML in Networks (clause 3) –** applicable datasets for the use cases identified in clause 2
- **AI/ML models for Networks (clause 4) –** relevant AI/ML models applicable for the use cases and datasets in clauses 2 and 3

  Future studies could include:

  - consideration of pipeline aspects which could be standardized for the different use case domains, the metadata for the model and the model context, e.g., the type and placement of the model with respect to ITU-T Y.3172 network levels, the training type of the model, the management techniques used for the model (Serving, Optimization, Training, Federation), the performance of the model with respect to the use cases and datasets (if available).

  - consideration of the different models approaches with their pros and cons in general. The different model approaches could discuss different architectural options (e.g., handling of the models by the operator versus by third parties, and their relevance in terms of data privacy and model performances) and these aspects could be also re-addressed from relevant perspectives (handling of the data privacy) in clause 5.
- **Toolsets for AI/ML in Networks (clause 5) -** toolsets for the use cases, datasets and models analysis, together with guidelines for usage of toolsets with respect to specific categories of use cases

  Future studies could include additional data generation tools, labelling tools, testing and validation tools, APIs and hardware accelerators in addition to software toolsets.
- **Overall summary (clause 6)** – summary of the content presented in the TR, mapping use cases, datasets, AI/ML models and toolsets
- **Analysis and remarks (clause 7)** – only some general considerations in this First Edition

  Future studies could include lessons learned for the specific use cases as well as general lessons learned. It will be also for consideration to highlight the open issues for the different identified use cases (datasets, models, toolsets) and in general.
- **Recommendations for future standardization and research activities (clause 8) -** potential future steps, including in terms of additional CG activities, standardization and research work, and collaborations (recommendations to ITU-T and beyond). *NOTE - A future enhanced version of the TR is expected to be an in-depth product for effective standardization initiatives.*

Use cases summary – examples extracted from the CG TR

Use cases	**Clause 2.2 telecom network operations supported by LLMs** **Examples includes:** **1. Network Anomalies Resolution** **2. Standard documents comprehension for engineer operations support** **3. Network Modelling for AI based optimization**
Datasets	Examples described in clause 3.2 include: 1. TeleQnA 2. SPEC5G 3. NetEval 4. CyberMetric:
Models	Clause 4.2: Large Language Models
Toolsets	Clause 5-2: 4 steps process: 1) Definition of the expected knowledge by a telecom Foundational LLM model, 2) Knowledge evaluation methodology, 3) Questionnaire creation, and 4) Questionnaire validation.

Table 7 Summary of the use cases for telecom network operations supported by LLMs

Use cases	**Clause 2.3 Autonomous Networks** **Examples includes:** **1. Import and export of knowledge in an autonomous network** **2. Configuring and driving simulators from autonomous components in the network** **3. Peer-in-loop** **4. Configuring and driving automation loops from autonomous components in the network** **5. Domain analytics services for E2E service management** **6. Automation and intelligent operation, maintenance and management (OAM) of radio network**
Datasets	Examples include: 1. GNN dataset in AI/ML Challenge 2. The openRan Gym as described in FGAN-I-329-R1 3. The FG AN Build-a-thon Argilla workspace
Models	Examples include: 1. GNN for inferring the network performance 2. LLM used for generating the experimentation scenarios and simulator configurations 3. An annotation model 4. A generative model trained for generating a complex composition of the containers. 5. A generative model trained for chaining a set of operations provided by the service frameworks.
Toolsets	FFS

Table 8 Summary of the use cases for Autonomous Networks

Conclusion – with thanks to all speakers and participants

The ITU CG-datasets will present its Final Report at the July 2024 SG13 meeting during its mid-plenary on 22 July for consideration and any future follow up - the Final Report will include a request of extension of the CG life time of one year

Today’s workshop is an additional great opportunity to exchange views on these AI matters and identify potential (research and) standardization activities as suggestions for ITU-T (SG13)

We wish to thank all participants in the workshop – and obviously all the speakers for their kind collaboration – and we hope all of participants will find the workshop interesting

CG-datasets access link:

https://www.itu.int/en/ITU-T/studygroups/2022-2024/13/Pages/Correspondence-Groups.aspx