联合国 E/CN.3/2019/27

经济及社会理事会

Distr.: General
19 December 2018
Chinese
Original: English

统计委员会

第五十届会议

2019 年 3 月 5 日至 8 日

临时议程* 项目 4(e)

供参考的项目：官方统计使用大数据问题

官方统计使用大数据问题全球工作组的报告

秘书长的说明

依照经济及社会理事会第 2018/227 号决定及以往惯例，秘书长谨转递官方统计使用大数据问题全球工作组的报告。报告中，全球工作组回应了统计委员会上届会议提出的要求，特别是在官方统计使用可信数据、可信方法和可信学习问题“联合国全球平台”上为全球统计界提供产品和服务，解决关于隐私和保密的关切，并提供有关全球平台业务模式的更多详情。全球平台是共享和测试可信方法、可信数据、可信培训材料的研究和开发环境，协同私营部门、学术界和民间社会为官方统计工作提供技术基础设施和服务。此外，全球工作组下属各任务小组(分别涉及地球观测数据、移动电话数据、社交媒体数据和扫描仪数据以及隐私保护技术)还举办了培训讲习班，编制了手册，并合作实施了创新项目。现请统计委员会注意该报告。

* E/CN.3/2019/1。

官方统计使用大数据问题全球工作组的报告

一. 导言

1. 统计委员会在 2014 年第四十五届会议上设立了官方统计使用大数据问题全球工作组。全球工作组负责按照职权范围(见 E/CN.3/2015/4)和统计委员会第 46/101 号决定(见 E/2015/24-E/CN.3/2015/40)，就官方统计(包括汇编《2030 年可持续发展议程》可持续发展目标各项指标落实情况)使用大数据全球方案制定战略愿景、提供指导并进行协调。

2. 统计委员会在其第 49/107 号决定(见 E/2018/24–E/CN.3/2018/37)中重申，使用大数据和其他新数据源对于国家统计机构实现现代化以紧跟数据环境快速变化的步伐至关重要，并突出强调了当前使用大数据填补空白、提升统计业务成本效益水平、取代调查并提高产出细化程度的机会。在该决定中，委员会核可了全球工作组关于进一步开发一个全球平台并以之作为协作型可信数据、可信方法和可信学习研究和开发环境的提案，重申需要提出关于该平台的企划案，鼓励全球工作组在迄今已取得的成果的基础上，为全球统计系统提供实用的产品和服务，以支持制作各项统计数据和指标，包括可持续发展目标各项指标，并强调必须谨慎应对数据信托、伦理、隐私、保密和安全所带来的社会挑战。

3. 本报告第二节着重介绍了 2018 年 10 月 21 日在迪拜举行的全球工作组年度会议和开放日。开放日活动包括官方统计使用可信数据、可信方法和可信学习问题全球平台会议和全球工作组下属各任务小组工作会议。[1] 第三节进一步详细介绍了联合国全球平台以及全球工作组下属各任务小组取得的工作成果，包括能力建设活动。第四节概述了全球工作组为推进落实其工作方案拟采取的下一步措施。

二. 全球工作组年度会议和“开放日”

4. 全球工作组第五次年度会议的会期比往年要短，这是因为会议当天恰逢“联合国全球平台开放日”。虽是如此，年度会议仍涵盖了下述议题：计划于 2019 年春季在基加利召开的大数据问题第五次国际会议的组织筹备；在联合国全球平台方面取得的进展；有待全球工作组处理的数据管理问题，特别是全球平台管理问题；全球工作组下属各任务小组工作成果简述；以及待提交统计委员会的全球工作组报告的编制。

5. 全球工作组往年都是在其大数据问题国际会议之前举行为期一整天的年度会议，2014 年北京会议、2015 年阿布扎比会议、2016 年都柏林会议以及上一次的 2017 年波哥大会议皆是如此。关于这些会议的报告以及全球工作组主席团会议的报告和文件，目前均可在全球工作组网站上查阅。[2] 全球工作组第五次年度会议原本计划依照惯例在大数据问题第五次国际会议之前举行，而按照传统时间

[1] 会议报告见全球工作组网站，可登录 https://unstats.un.org/bigdata/查阅。

[2] 见 https://unstats.un.org/bigdata/bureau/。

表，该次国际会议将在 2018 年秋季召开。但在关于第二届联合国全球数据论坛定于 2018 年 10 月底在阿拉伯联合酋长国迪拜举行的通知发布之后，全球工作组主席团立即决定将大数据问题国际会议推迟至2019年春季在卢旺达基加利举行。而另一方面，全球工作组年度会议又必须举行，以避免出现连续 12 个月不举行全体成员实体会议的情况，因此破例安排在联合国全球数据论坛开幕前一天举行。

6. 全球工作组的两大关注点分别是联合国全球平台和相应的数据管理问题。全球平台已从全球工作组的概念落地为现实，开始交付数据、方法和学习。在这方面，全球工作组尚待更精确地界定并商定其四个基本支柱的概念：可信数据、可信方法、可信伙伴和可信学习。这就意味着要商定全球平台上各类大型数据集的所有权和访问权限、数据或算法是否需要“公开”以及软件、服务和工具如何实现“独立于平台”。这些问题直接关乎全球平台的业务模式。对此，全球工作组决定分别与大不列颠及北爱尔兰联合王国国家统计局和加拿大统计局紧密合作，编写一份关于业务模式的文件和一份关于数据管理的文件。

7. “全球平台开放日”由全球工作组组织，由阿拉伯联合酋长国联邦竞争力和统计局主办。“开放日”方案包括全球平台展示以及其后举行的关于农作物和土地覆被统计(使用卫星数据)、关于测量人口流动情况(使用移动电话数据)、关于测量价格波动情况(使用扫描仪数据)和关于隐私保护技术的会议。[3]。

8. 全球平台对全球统计界数据项目合作保持开放。平台所收录的数据集日渐增多，其中包括大地卫星和哨兵卫星数据、来自行星实验室公司网站(Planet.com)的试验卫星数据、通用船载自动识别系统船舶定位数据以及广播式自动相关监视系统航空器定位数据。平台可提供包括各类云服务器、地球空间分析服务、Jupyter 笔记本在内的多种服务。联合王国国家统计局数据科学学院利用全球平台开展联合王国 112 座城市城区植被指数(利用来自谷歌街景的 1 700 万张图像)等研究或估算北爱尔兰居住在四季通行的道路两公里之内的农村人口所占比例(可持续发展目标指标 9.1.1)(使用开放街道地图和人口数据)。

9. 随后举行的“开放日”会议上展示了新数据源和新技术在官方统计工作中的应用，其中部分官方统计还一直在利用全球平台实施项目。随着时间的推移，全球工作组希望能有更多项目积极利用全球平台，并存储经过测试的方法和数据，以供对外分享使用。哥伦比亚国家统计局曾在某试点项目中利用一个内含卫星图像处理结果和其他数据源的确定性模型来估算谷类作物产量。除其他外，哥伦比亚国家统计局还利用谷歌地球软件预处理和处理卫星图像的算法来获得估算作物产量所必需的条件植被指数和条件温度指数。

10. 加拿大统计局借助卫星数据来改进作物产量模型。该统计局使用卫星数据测试(自 9 月份开始)19 种作物的产量估计数，经测评，其中 15 种作物的产量估计数精度之高足可作为官方统计数据公布。这就意味着作物产量调查可逐渐为基于卫星数据的作物产量估算所取代。总而言之，加拿大统计局得出结论认为，必须

[3] 详情可查阅 https://unstats.un.org/unsd/bigdata/conferences/2018/open-day/default.asp“议程”项下。

加快学习，为进行试验和评估质量创造有利环境。全球平台所能提供的惠益可能就在于利用可信方法、可信数据和可信伙伴关系为合作提供便利。

11. 联合国环境规划署(环境署)演示了“全球表层水探索者”应用程序，该款应用程序为免费和开放获取关于水域范围的国家、流域和次流域合并数据开辟了一条途径，可为衡量可持续发展目标指标 6.6.1 的落实情况提供支持。对尚不存在其他数据源的领域而言，这还有鼓励扩大参与之效。此外，该全球应用程序确保了跨越时空的数据对比性，并提供了一个时间回溯窗口。该款应用程序充分利用了全世界地球观测算法的专业知识精髓。

12. 数据分析公司 Positium 助力印度尼西亚统计局改进旅游统计质量，利用移动定位数据填补了邻国游客跨境统计的空白。此外，利用移动定位数据可以更准确地估算境内旅游出行人次和目的地数据，同时还能提高统计频率，降低统计成本。

13. 欧盟统计局目前正在构建关于处理移动电话数据以供官方统计之用的统一方法论。这个俗称的参考性方法论框架将促进移动网络运营商与统计人员在技术和组织层面互通合作，并确保处理方法具有一致性、可重复性和可接续性。此种合作还将为明确《通用数据保护条例》等法律问题奠定扎实基础，而且有助于实现对多个移动网络运营商的同步处理和分析(即融合来自不同运营商的数据)。

14. 欧盟统计局提议从三个层面开展工作：一是处理原始数据的移动网络运营商数据层；二是应用和处理统计方法的统计层；三是使用统计方法加工机密移动电话数据的中间汇聚层次。在汇聚层，统计人员可以设计算法，移动网络运营商可以利用安全多方计算技术执行这些算法。

15. 国际移民处在政治辩论的风口浪尖之上，因此，对移民统计的需求之迫切前所未有。秘书处经济和社会事务部统计司支持各国努力加强自身收集和传播移民存量和流动统计数据的能力，并与多个项目国合作，其中包括格鲁吉亚，该国现已专门成立国家移民问题委员会，初步使用普查、调查和行政数据建设综合性国家移民数据基础设施。这项工作可助力落实与按移民身份分列的各个细目有关的所有可持续发展目标指标。

16. 除了上述更传统的数据，该项目还开始研究使用新的数据源和技术，并与格鲁吉亚国家移动网络监管机构建立了合作伙伴关系，该监管机构有权访问和处理移动电话数据。统计司与欧盟统计局、国际电信联盟、国际移民组织和 Positium 等国际伙伴合作，尝试利用移动定位数据和社交媒体数据改进人口流动情况测量工作，确定移民、旅游和通勤者统计数据。

17. 到目前为止，实现价格指数测算现代化的努力侧重于使用来自零售商的扫描仪数据，目的是提高扫描仪数据在官方统计中的有效使用水平。全球工作组扫描仪数据任务小组想交付一项可挂靠在全球平台上的工具，使用信息、数据和测算公司尼尔森的历史扫描仪数据进行分析、监测和指数估算，同时提供关于该工具使用方法的培训和说明材料、方法指导材料以及良好做法目录。除其他外，扫描仪数据为随着时间的推移逐渐改变支出权重提供了机会。目前，各国国家统计局

在全球平台上既可获取经过测试的指数方法代码，还能以某些培训数据为素材，使用不同方法进行指数计算练习。

18. 尼尔森现已成为全球工作组合作伙伴。对尼尔森而言，数据即是其业务，该公司负责收集、扩增和提供与人们所看、所听和所买之物有关的数据。尼尔森还已着手在全球各地收集电子商务数据，截至 2018 年，其数据收集工作已覆盖全球 76%的区域。过去一年里，尼尔森和全球工作组一直在寻求利益共同点，目前双方合作迎来双赢局面。

19. 2018 年初，全球工作组成立了隐私保护技术任务小组，由联合王国国家统计局领导，按照预想，其任务是起草治理和信息管理数据政策框架中的加密部分。该任务小组将制定和提出全球平台内部加密原则、政策和开放标准。这当中将涉及数据使用的伦理问题，在制定数据收集、处理、存储和列报方法及流程时充分考虑到数据隐私、保密和安全问题。除其他外，这将减少专利信息和敏感信息处理过程中可能发生的风险。

20. 信通技术公司 Cybernetica 作为全球工作组隐私保护技术任务小组的正式成员，演示了隐私保护技术对扫描仪数据和移动电话数据的应用。密码术可用来阻断同一对象再次识别的可行性，在不降低结果准确度的情况下减少内部攻击风险。相关例子包括安全多方计算和同型加密，这两种技术已在 Cybernetica 与爱沙尼亚政府的合作案例中得到成功运用。此外，还可采用匿名化技术来添加噪声，从而增加数据对象再次识别的难度，但这类技术也会降低结果的准确度。差别化隐私和 k-匿名化就属于此类。

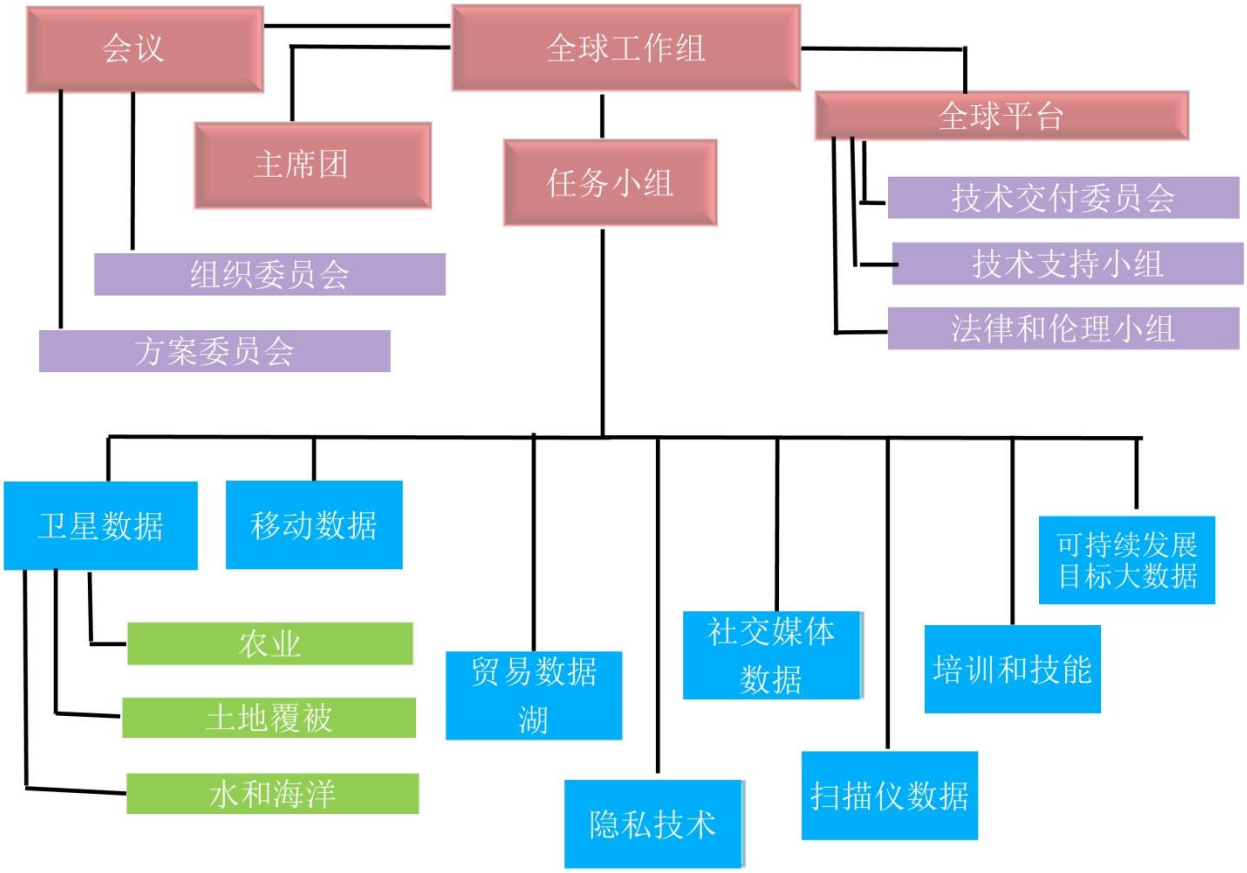
三. 联合国全球平台和全球工作组下属各任务小组

21. 全球工作组成立四年以来，一直通过下属的几个任务小组、全球平台委员会、国际会议组织委员会以及相关工作流和项目主管委员会组织开展活动。全球工作组主席团每两周举行一次会议，以管理和指导各任务小组和委员会的工作。各任务小组和委员会分别在牵头机构的领导下开展工作，由牵头机构自行安排定期会议和决断工作方案。全球工作组的详细组织结构见下图。

A. 联合国全球平台

22. 《波哥大宣言》(见 E/CN.3/2018/8)阐述了全球平台作为共享、交换和开发可信数据和元数据、可信方法、可信伙伴和可信学习之促进者的详细情况和驱动因素。全球平台将加入相互联结的联合平台网络，以公私大数据倡议的最佳做法为基础，为整个官方统计界进行数据创新提供技术基础设施。按照预想，全球平台将透过可信培训材料、方法和软件应用程序图书馆，通过举办关于官方统计现代化、使用替代数据源(例如大数据)以及新工具、服务和分析技术应用的讲习班，支助开展能力建设。全球平台的开发和维护工作将在统计委员会主持和指导下进行，为发达国家和发展中国家的国家统计系统提供支持。

图
全球工作组的组织结构



23. 全球平台现可提供若干阿尔法服务，例如访问阿里巴巴云、亚马逊网络服务、谷歌云平台和微软蔚蓝云，以及代码协作、方法发布和地球观测以及位置数据分析等众多其他服务。全球平台的用户可以搜索、构建、部署和使用算法和统计方法，并且可以利用统计界使用的主要编程语言(R、Python、Java 和 Scala)进一步开发方法。全球平台还可托管机器学习模型并向模型发布应用程序接口端点。分布在世界各地的全球平台合作伙伴可以通过调用应用程序接口使用自身所处环境中的算法，而且今后还能够访问多个全球数据集，例如最早可追溯至 2016 年 7 月的广播式自动相关监视系统飞行数据、通用船载自动识别系统船舶数据以及高分辨率商用卫星图像。

B. 全球工作组下属各任务小组

24. 卫星图像和地理空间数据任务小组[4] 在交付手册一年之后，又成功举办了一次为期五天的培训讲习班，内容涉及使用地球观测数据进行统计以及使用卫星图像数据估算作物产量和进行相关统计的方法。由于增列了不属于农作物统计领域

[4] 见 https://unstats.un.org/bigdata/taskteams/satellite/。

的议题，该任务小组特意建立了三条工作流，分别是农作物产量估算、土地覆被和土地使用统计以及与水有关的生态系统统计。

25. 在农业统计工作流中，加拿大统计局出于共享和测试目的，向全球平台上传了自有卫星数据、作物调查数据、农业气候数据、部分加拿大作物产量源代码以及有关文件。这种协作型共享和测试做法可最终推动建立可信数据、方法和学习。该条工作流将有助于为衡量可持续发展目标指标 2.4.1“实行生产性和可持续农业做法的农业地区比例”落实情况提供参考信息。另据估计，加拿大、卢旺达和肯尼亚还将实施若干利用卫星数据识别具体作物并判断其产量的定点项目。

26. 土地覆被和土地使用统计工作流具体关乎可持续发展目标指标 11.3.1“土地消耗率与人口增长率的比率”和指标 15.1.1“森林面积占土地总面积的比例”。这两项指标包括测量土地覆被和土地使用随时间推移的变化和状况。测量所用方法经联合国环境经济核算专家委员会确认有效。该条工作流着眼于实施一个测量泥炭地随时间变化情况的项目。

27. 卫星图像和地理空间数据任务小组第三条工作流重点关注与水有关的生态系统范围随时间的变化情况(可持续发展目标指标 6.6.1)。如前所述，现有一个全球应用程序可供利用卫星数据估算淡水范围。在国家一级，可根据当地具体情况对该全球应用程序略作调整，例如，加拿大方面考虑到该国大部分地区的水面在冬季几个月份会结冰，就对该程序作了调整。此条工作流目前正考虑在一个或多个发展中国家实施一个关于具体水流域的项目。

28. 关于全球工作组下属的其他任务小组，使用移动电话数据任务小组手册的草案初稿已全部完成，其中详细介绍了主要用于估算旅游统计数据的数据源、方法、伙伴关系模式和应用程序。该手册第二卷的编写工作即将展开，该卷内容将包括可供测量人口流动情况的其他应用程序，还将涉及通过设计保护隐私的问题，例如前文所述的欧盟统计局框架。该任务小组高度重视测量移民、旅游及相关统计数据项目在格鲁吉亚的实施工作，而且计划在哥伦比亚、印度尼西亚和意大利测试在该项目实施期间制定的方法和算法。

29. 使用扫描仪数据计算居民消费价格指数任务小组目前正在以开源应用程序为主要工具，测试统计方法和软件代码，并且已在某手册中对测试情况加以记录。这样，其他统计局便可在此基础上就扫描仪数据以及网络抓取数据和调查数据在各自统计资料编制工作中的潜在使用进行试验和测试。另外一项可喜的发展是与尼尔森公司建立了合作关系，该公司现已将某些数据向全球统计界开放。尼尔森数据作为一个数据源所提供的数据具有全球性和标准化特点，因而也具有可比性，这可使计算居民消费价格指数的可信方法得到共享。

30. 最后，正如全球工作组开放日期间所介绍的那样，全球工作组新成立了一个关于隐私保护技术的任务小组，负责制定隐私保护方法和程序，以确保可信伙伴在全球工作组平台上处理和交流专利信息和敏感信息时无安全之虞。该任务小组将使用开放标准和开源算法，制定和提出加密原则和政策。目前，该任务小组正在编写一本手册，预计该手册将在 2019 年面世。

四. 下一步措施

31. 全球工作组需要进一步发展全球平台作为全球统计界协作研究和开发环境的可持续业务模式。项目和可信学习是全球平台相关性和可持续性概念的证明。2019 年，全球工作组将组织和参与一系列活动，以展示全球平台的进展和就绪情况。

A. 联合国全球平台的业务模式

32. 全球平台业务模式的主要方面具体如下。更多详细信息将在本报告的背景文件中加以介绍。

法律实体

33. 法律实体的必要性在于充当全球平台整体运作的载体，因此必须明确该实体的架构和所有权。按照预想，该实体将负责筹集和管理资金，并且有能力为与全球平台运行有关的风险做担保。诸如环境署世界养护监测中心 等模式或可提供良好的解决方案。

业务实体

34. 按照设想，全球平台的可信伙伴既有产品和服务的提供者，也有产品和服务的用户。同一参与机构的不同部门可能既是提供者又是用户，其中，官方统计机构及其合作伙伴具有特殊地位。需要针对不同类型的合作伙伴确立参与规则。全球平台在初期将侧重于开放数据源，允许随着时间的推移对更敏感的数据进行差异化访问。全球平台旨在向全球提供全天候支持。此模式或将推广至区域枢纽，促进共同开发和能力建设活动。

供资

35. 一旦获得可观的资金，全球平台便可迅速发展壮大，而所需资金预计将来自开发和慈善两个来源。典型的投资者可能包括发展资金来源、基金会或大型技术提供者捐赠的慈善资金。

B. 概念证明：项目

36. 目前，几乎每一个任务小组都已着手就须作为概念证明在全球平台上实施的某一个或某几个项目展开工作。这些项目处于不同的发展阶段，其中，有必要着重指出以下项目：

(a) 使用卫星数据估算作物产量。该项目的一个良好实例是上文提到的加拿大统计局所开展的工作，即成功地利用卫星数据估算出了 15 种作物的产量。还计划针对非洲国家实施类似的项目；

(b) 测量土地覆被和土地使用的变化情况。计划实施一个重点测量泥炭地随时间变化情况的项目；

(c) 测量与水有关的生态系统的范围。当前正在开展的一个项目是估算加拿大淡水范围，该项目可能还会在具体的三角洲区域(例如湄公河三角洲)实施；

(d) 测量人口流动情况。现已在格鲁吉亚启动一个项目，利用移动电话数据测量移民、旅游、季节工人和白天/夜间人口情况；

(e) 使用尼尔森数据估算居民消费价格指数。加拿大现已完成使用尼尔森数据估算价格指数的初步测试工作。2019 年计划在更多国家对更多产品展开进一步测试；

(f) 使用贸易和交通数据进行探索性数据分析。现已在全球平台上构建一个数据湖，其中载有贸易统计数据、通用船载自动识别系统船舶数据和广播式自动相关监视系统飞行数据，数据测试将在 2019 年进行。

C. 概念证明：可信学习

37. 2017 年 11 月在波哥大举办了 4 次为期两天的培训讲习班，这表明全球工作组完全有能力开设关于使用新数据源汇编官方统计数据的课程。2018 年 6 月在曼谷针对亚洲及太平洋区域举办了一次为期5天的使用卫星图像数据进行作物及相关统计区域培训讲习班。2019 年计划继续举办类似的培训讲习班，包括在大数据问题第五次国际会议召开之前举办多场为期较短的培训讲习班，以及于 2019 年6月上半月在雅加达举办一次为期五天的使用移动电话数据进行官方统计培训讲习班。

38. 培训、技能和能力建设任务小组现已开启工作新阶段，出现了新的可交付成果，在这一阶段，任务小组更多地侧重于在不断变化的数据环境中发展技能。波兰统计局在这方面一马当先。让整个欧洲统计系统拥有大量掌握数据科学技能的统计专业毕业生是欧洲的一个长期目标。到 2020 年，数据科学技能将成为官方统计教育不可或缺的组成部分。全球工作组技能发展方案将与世界各地的现有方案，特别是欧洲统计培训方案紧密联系起来。欧洲统计培训方案一直在开设培训班，教授如何使用大数据源，例如社交媒体文本分析和网络搜索或使用移动电话数据进行官方统计。

D. 活动

39. 2019 年，全球工作组将有多个机会展示全球平台的进展。全球工作组将组织开展以下活动：

(a) 定于 2019 年 3 月在纽约举行的统计委员会第五十届会议会外活动。如前所述，届时，全球工作组将提交本报告的背景文件，其中将列明全球平台可持续业务模式备选方案。会外活动期间，这些备选方案在稍作解释后，将开放供更广泛的统计界讨论；

(b) 大数据问题第五次国际会议。继亚洲及太平洋、中东、欧洲和南美洲之后，非洲将主办这次大数据问题全球会议，届时，非洲大陆上的国家将有机会展示新数据源和技术在各自官方统计数据汇编工作中的使用情况。如上文已经指出

的那样，将在国际会议间隙组织举办培训讲习班，广泛涉及各类与大数据有关的主题；

(c) 定于 2019 年 8 月 15 日至 17 日在吉隆坡举行的统计学会第六十二届世界统计大会关于大数据和新技术的卫星活动。全球工作组与世界统计大会组织委员会合作，计划在全球平台上提供卫星数据、移动定位(培训)数据、社交媒体(培训)数据、扫描仪(培训)数据使用实训，目标受众是来自各统计局的统计人员和数据科学家。

五. 需要统计委员会采取的行动

40. 请统计委员会注意本报告。

附件

官方统计使用大数据问题全球工作组成员

国家

澳大利亚

孟加拉国

巴西

喀麦隆

加拿大

中国

哥伦比亚

丹麦

埃及

格鲁吉亚

德国

印度尼西亚

爱尔兰

意大利

墨西哥

摩洛哥

荷兰

阿曼

巴基斯坦

菲律宾

波兰

大韩民国

沙特阿拉伯

瑞士

阿拉伯联合酋长国

大不列颠及北爱尔兰联合王国

坦桑尼亚联合共和国

美利坚合众国

组织

非洲开发银行

加勒比共同体

亚洲及太平洋经济社会委员会

非洲经济委员会

欧洲经济委员会

欧盟统计局

联合国粮食及农业组织

国际货币基金组织

国际电信联盟

经济合作与发展组织

海湾阿拉伯国家合作委员会统计中心

亚洲及太平洋统计研究所

统计司

联合国“全球脉搏”举措

万国邮政联盟

世界银行