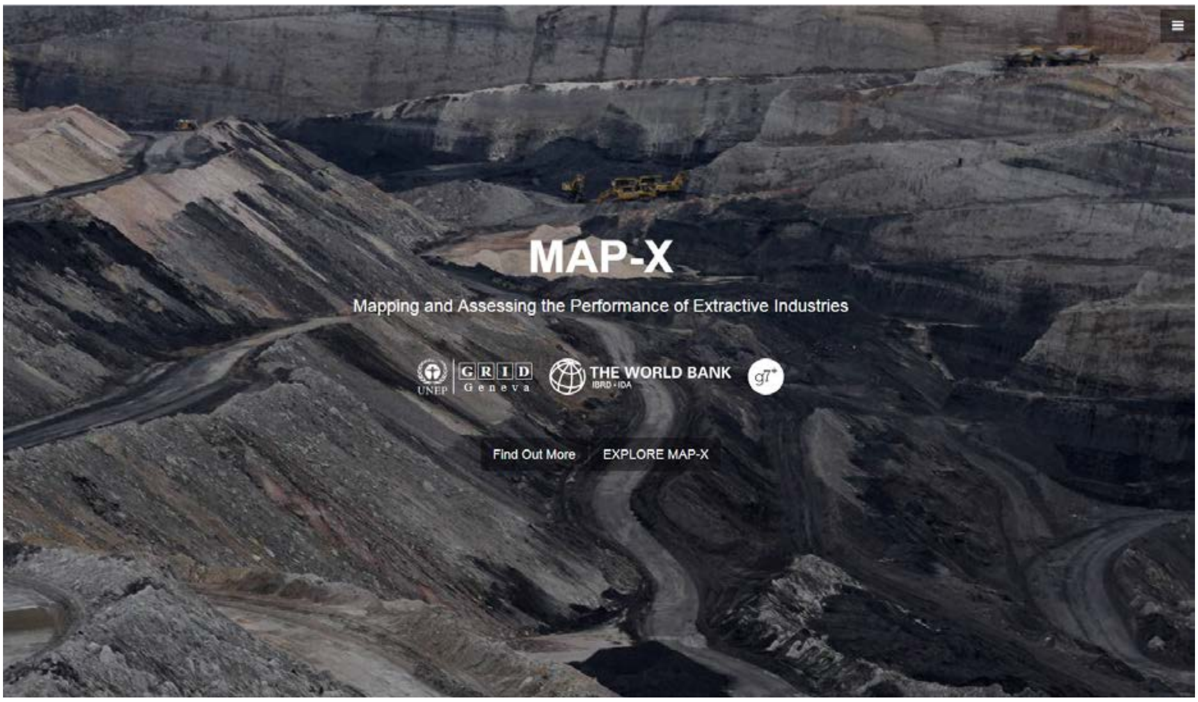
CERTIFICAT COMPLÉMENTAIRE EN GÉOMATIQUE

Consolidation of the open source platform MapX for the extractive sector



REPORT

CANDIDATE: ANTONIO BENVENUTI

SUPERVISOR: DR. PIERRE LACROIX

October 2017

GRID, International Environment House, 11 chemin des Anémones 1219 Châtelaine, Genève (Suisse)

Acknowledgments

My largest acknowledgments go to Dr. Pierre Lacroix who trusted in me, supervised me, and gave me the possibility to be an active part in the development of such an important project, at its most interesting stages, that is MapX. Secondarily I kindly thank Dr. Nicolas Ray for his support on my cause and his kindness. If I had a very good time working at GRID is also because of the people that work there. I am grateful to have worked with you for these three and a half months. Especially, I would like to thank all the persons that were closer to me, namely Thomas, Fred, Alejandra, Laetitia, Audrey, Lorenzo and Erica.

Table of contents

Annexes (Internal documents that were produced during this internship and that can be accessed in the Owncloud by GRID users)

Annexe 1: *Key-functions.docx*

This document contains an introduction to the functionalities of MapX v3. It was conceived as didactic material for teachers that will introduce MapX to future users at the country's scale.

Annexe 2: *Github guidelines.docx*

This document explains in detail the steps for reporting bugs and proposing new features in the MapX application. It applies to stakeholders at country's scale.

Annexe 3: *Reporting activities 2.4 & 2.10_AB.docx*

This is the final report of the Afghanistan project for the activities that were developed during this internship.

Annexe 4: Report assessment DRC.docx

This document contains a detailed report on the assessment of feasibility for the inclusion of financial data published by the Extractive Industries Transparency Initiatives of the Democratic Republic of Congo for the years 2007 to 2012.

Annexe 5: *tracking reports.xlsx*

This spreadsheet is a complement file for the report assessment DRC.docx. It contains all the relevant information that must be used when approaching the creation of the EITI database to be published in MapX.

Annexe 6: *Layers Afghanistan.docx*

This document contains the list and categorization of all available layers in MapX for the Afghanistan project and those that were selected for WMS streaming from the iMMAP GeoNode.

Table of figures

Table of tables

Summary

This internship was performed at UN Environment – GRID in Geneva in the frame of the Complementary certificate on Geomatics for three and a half months, in the period of July – October 2017. It focused on spatial data infrastructures and geoprocessing and was supervised by Dr. Pierre Lacroix (jury member together with Dr. Nicolas Ray). In particular, this work was committed to the consolidation and testing of the platform MapX for spatial data and the evaluation of its level of interoperability with the potential and effective data providers. Since Afghanistan's adherence to the MapX project in early 2017, a large amount of data concerning the extractive sector and related domains that was received from UN Environment Afghanistan had to be screened, processed and published into MapX. This process was pursued following a mechanism of rectification, validation and a normative standard. Missing data were collected from web sources from which data of different levels of compliance to the MapX standard could be obtained. Geoprocessings and data preparation in general was accomplished on the ArcGIS and QGIS environments when needed. The evaluation of the level of interoperability between MapX and the other platforms allowed assessing the potential problematics that should be solved before MapX could establish as the reference platform for spatial data on natural resources. In order to reduce the gap between MapX and the external data platforms, and to facilitate the publication of a large number of time-related datasets in MapX, a Python script for time-values transformation into the "POSIX" format was conceived and redacted. Furthermore, this project focused on the process of preparation of non-machine readable time-series data to be published in MapX. Specifically, an assessment of feasibility was performed on the Extractive Industry Transparency Initiative (EITI) data for the Democratic Republic of Congo (DRC) in the annual reports of the years 2007 to 2012. The intensive use of the MapX application made possible contributing to its growth with stability and user-experience testing. Accordingly, bugs and requests of new functionalities for the application were proposed via the Github platform. This internship finally resulted in the production of teaching material to instruct thirds on how to share and teach MapX to universities and institutions at the country's scale. For the sake of the Afghanistan project and to expand its action to other contributors in Afghanistan, this material was used for a teaching session to colleagues from UN Environment Afghanistan.

1. Introduction

1.1. Host: UN Environment - GRID

The mission of the UN Environment is to provide leadership and encourage partnership in caring for the environment by inspiring, informing, and enabling nations and peoples to improve their quality of life without compromising that of future generations (UN Environment). To achieve this mission, the United Nations programme works on three main axes:

(1) Assessing global, regional and national environmental conditions and trends;

(2) Developing international and national environmental instruments;

(3) Strengthening institutions for the wise management of the environment.

The Global Resource Information Database (GRID) network was created in 1985 as a response to these three axes of work. This office, with headquarters in Nairobi – Kenya, is a global network with 15 centres managed by the Science Division of the UN Environment. The GRID's main objectives are:

• facilitate access and produce high-quality environmental data and information to support decision-making and policy-making;

• support the United Nations Environment Programme in its ongoing review of the state of the environment and trends;

• provide pre-alerts on emerging environmental issues and threats.

The GRID-Geneva is one of the two oldest centres of the GRID-network. It is composed of three institutional partners: UN Environment, the Federal Office for the Environment (FOEN) and the University of Geneva (UNIGE), which have been linked through a multi-annual partnership agreement since 1998. GRID-Geneva is a key centre of geo-spatial know-how, with strengths in GIS, IP/remote sensing and statistical analyses, integrated through modern spatial data infrastructure and web applications (UN Environment-GRID).

1.2. The "MapX" Project

MapX (available at http://app.mapx.org) is an open access geo-mapping platform for extractive industries in fragile states that is developed by UN Environment in partnership with the World Bank. In line with its initial name (MAP-X → Mapping and Assessing the Performance of Extractive Industries), the platform was created to assist stakeholders in the extractive sector through publication of reliable data that directly and indirectly concern the extractive domain. Recently, the platform has evolved into a wider tool for consolidating all existing resource concession, land use, and risk information into a single open source system and providing transparent access to this information by all stakeholders, improving strategic decision making and enhancing benefit sharing.

MapX was designed to address the problems caused by fragmented, inaccessible, out-dated or contested data thus creating prerequisites for building trust and sharing benefits. The innovative technology platform aggregates, authenticates, and publishes geo-spatial data on the combined impacts of extractive projects. By providing open access to impartial and authoritative data, MapX aims to address information asymmetries and improve the quality of dialogue between stakeholders. MapX will build the capacity of stakeholders to effectively understand and use the resulting information to identify key risks and benefits, shape the content of benefit sharing discussions, and to directly participate in the monitoring of compliance and performance at the concession level. The platform aims to improve transparency, help stakeholders understand the distribution of costs and benefits across the project life cycle and, as a result, build trust and reduce the potential for conflict.

The possibility of offering MapX as an open platform of data that could support the different initiatives on extractive industry transparency (EITI) and position/embed it as an open data tool for the Petroleum (MoMP)/World Bank and the National Advisory Board (donors, government, non-government) requires supporting the establishment of MapX as a reference platform for each country. MapX has been tested in several countries such as Afghanistan, the Democratic Republic of Congo (DRC) aligning closely with the EITI process, Nigeria, Colombia, South Sudan, Somalia, Haiti, and at global scale for the Minamata Project (on environmental pollution caused by mercury). These projects demonstrated the potential value that MapX can provide in terms of visualizing conflict dynamics around mining sites but it also showed the infinite multiple uses of this platform in terms of data interaction and visualization. For example, the Minamata project has demonstrated the power and the high configurability of the story map function for communication purpose in MapX. Thanks to the high flexibility demonstrated by MapX, the application has become the official data-visualization platform of GRID and will be implemented accordingly since 2018.

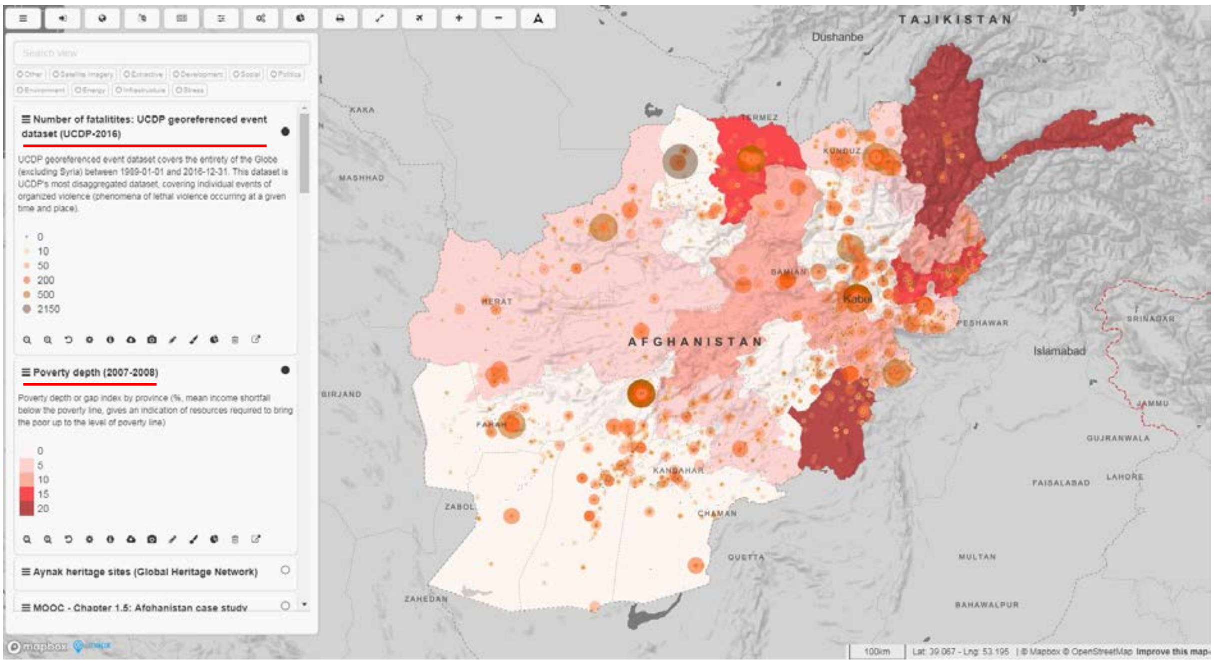


Figure 1: Visualization of conflict-related data against the poverty indicator in Afghanistan.

MapX key services

In synthesis, the MapX application offers the following key services to all stakeholders in the extractive sector:

- MapX consolidates and aggregates the various streams of financial, social and environmental performance information into a single on-line site to improve accessibility and reduce fragmentation (Figure 1).
- MapX authenticates this information against impartial and authoritative data quality benchmarks and provides stakeholders with a measure of trust (Figure 2).
- MapX provides users with a suite of analysis tools and dashboards to help visualize the information, understand the distribution of benefits and risks, as well as monitor contractual compliance and performance of specific projects over time and their respective contributions to specific Sustainable Development Goals (SDG)'s targets (Figure 3).

For a comprehensive description of the potential functionalities of MapX v3 read Annexe 1.

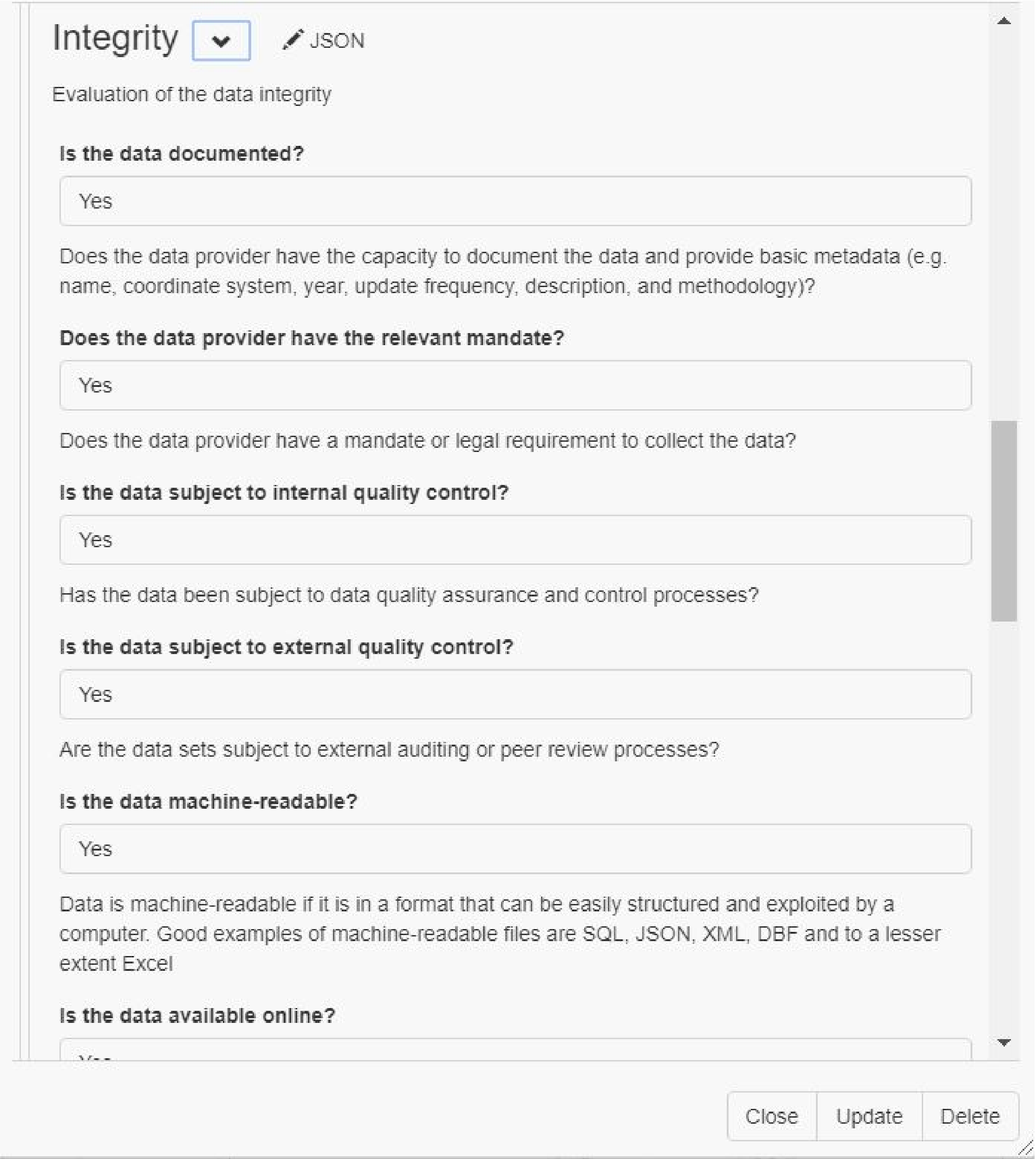


Figure 2: Assessment of data integrity and quality in the metadata of published layers in MapX.

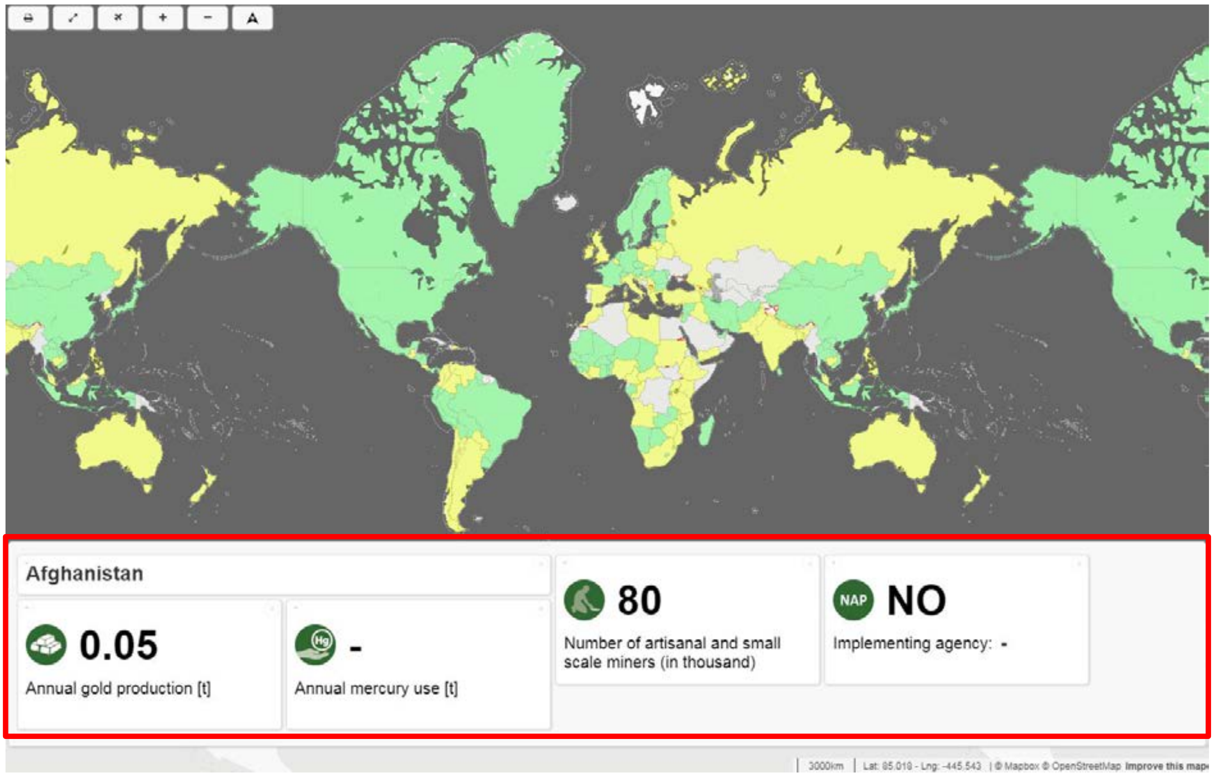


Figure 3: Dashboard available for a specific view for Afghanistan. The information in the dashboard complements the one available on map.

MapX data categories

- **Extractive data**: data that are directly or closely related to the extractive sector such as mining concessions, mineral deposits, areas of interest for mining activities, geological maps.
- **Financial data**: data on any natural resource concession-related payments made between different stakeholders (e.g. company-government, company-community, government-community payments). This can include data from EITI annual reports, revenue sharing, benefit sharing agreements and voluntary social payments.
- **Economic data**: economic development data including indicators on employment, infrastructure development, production data and local content (procurement of goods and services).
- **Social data**: social impact assessment data, information on human development indicators (education, poverty, and inequality data), distribution of indigenous populations, conflict data, armed groups, illegal activities, human rights, stress factors etc.
- **Environmental data**: environmental impact assessments, environmental baselines, and quality monitoring data (e.g. air, water, and soil quality), environmental performance targets at project closure, satellite images informing on land use and mining site extent, etc.

1.3. Aims of the internship

Consolidation of the MapX data platform

One of the most important goals of this internship was to consolidate the MapX platform with data that is directly or indirectly relevant for the extractive sector of developing countries. This process constitutes the basis for 1) demonstrating the potentiality of MapX on data visualization and interaction to the different stakeholders and 2) establishing MapX as a reference platform for the involved countries.

Data to be published were selected among different sources on the basis of relevance, quality and integrity. Before being published, data were edited in order to have a MapX -conform structure (e.g. right coordinate system (WGS-84), non-geometry-related attributes, spell mistakes in the attribute table, un-fragmented…) and to be as clean and integer as possible in order to be set available publicly.

This study largely focused on the Afghanistan project. The first part of this project kicked off early 2017 and will be concluded by December 2017. The relevant data of this project were in part transmitted by UN Environment Afghanistan in agreement with the local government. The rest of the data was sourced from web platforms.

Interoperability assessment between MapX and other data platforms

The assessment of the current level of interoperability between MapX and the different data providers represents another important goal of this study. The improvement of MapX and the creation of more suited data platforms from which data could be sourced and published into MapX constitute the basis for establishing MapX as the reference platform for the extractive sector at national levels.

In particular, this study focused on the data related to the Afghanistan project. Interoperability assessments were conducted between MapX and both specific and general data providers for Afghanistan. The government related sources that were evaluated were those from which a number of datasets received from UN Environment Afghanistan were sourced. Among those were the Afghanistan Central Statistics Office, the Afghanistan Information Management Services (AIMS), and the Afghanistan Spatial Data Center (iMMAP).

In addition to this, a study was conducted on the feasibility of publishing financial time series data transmitted by the Extractive Industries Transparency Initiative for the Democratic Republic of Congo (EITI-DRC) on the basis of their level of compliance to the MapX standard. The interoperability between the EITI database and MapX was investigated for the time period between 2007 and 2012.

Stability and User-experience testing

During March 2017, MapX has seen the launch of its third version of the application, which was made available for public in September 2017. During this intermediate period, the beta application was tested in order to report bugs and propose improvements while publishing relevant datasets for the MapX project.

Didactic: creating a guideline for the use of MapX v.3 and Github

Didactic material was prepared in order to give access to the functionalities of the application to users other than those at GRID-Geneva and to allow others to contribute introducing MapX to thirds. Specifically, the creation of a knowhow on MapX functionalities constitutes the basis for the spreading of MAP-X at the country's scale. In the frame of the Afghanistan project, three afghan partners came to Geneva to take part of a training program on GIS and MapX in begin September. For the sake of their course, I conceived - with Pierre Lacroix - and prepared the training material on major functionalities of MapX and taught it during a 1 afternoon course. The material is available in Owncloud as Annexe 1. Furthermore, as the partners became active contributors and users of MapX, it was given them the opportunity to contribute to the improvement of the application via the Github reporting platform. A guideline on how to use Github was also conceived and generated for this sake (Annexe 2).

2. Methodology

2.1. Consolidation of MapX data platform

The consolidation of the platform was based on the following steps:

1) Conceptualization:
   The conceptualization of relevant data that lack in MapX for each country was mostly conducted for Afghanistan.
2) Data seeking:
   For this project, roughly 30GB of spatial and non-spatial datasets were provided by UN Environment – Afghanistan. A process of screening of the material received was necessary in order to perform a gap-analysis and to select up-to-date and authoritative data to be published. Other data were sourced from web platforms (through direct download or through contacting a reference person) and others were selected from spatial data infrastructures (SDI). Large priority was given to data that featured metadata information.
3) Data preparation:
   Data was edited in order to be published in MapX through GIS software. ArcGis or QGis were used to perform different degrees of changes depending on the specific data status. Some data were simply checked for their integrity and coherency. Others had their attribute table edited. A few datasets were generated by combining two or more spatial datasets via vector geoprocessing. Data that were only available as Excel spreadsheets were first cleaned and prepared in order to be joined to a spatial dataset to which they could be associated. The data preparation included the conversion of "date"-type values into "posix" values in order to employ the spatial datasets in MapX as time-related data. This activity was accomplished by creating a Python script that simplifies the process of data conversion for a large number of datasets at the same time (Ch. 3.1 - Date to posix transformation)
4) Data publication:
   Data were uploaded into the MapX platform through its browser application (http://app.mapx.org). The system accepted both zip files containing shapefiles and geojsons.

Data were published as "source layers". Metadata were documented for each published source layer. The integrity of each layer was identified via a 22 steps questionnaire conceived by GRID-Geneva and documented as layer's metadata.

5) Views creation:
   Views were created from each published source layer for the concerned country/ies. A style based on a selected attribute was created through a selection of absolute attribute values and the associated colour, symbol, size.
   Each view was also characterized with the selection of supplementary attributes that could be queried on map when clicking on features in order to complement the available information given through the styling.
6) Tracking the uploaded source layers and views:
   A spreadsheet ideated by Alejandra Arango (previous intern at GRID) was used in order to keep track of all updated source layers and the views that were created out of them for each country. Relevant information included the editing done for the preparation of the layer, the missing metadata information and the unique name of the same layer in the back-up storage facility located in Owncloud. The same process was applied to the data for what concerns the integrity assessment of each source layer. The information was backed-up in country-specific spreadsheets where the integrity of the source layer was expresses though a traffic-light colour-scale. In the same spreadsheet, the authoritative source of each layer was defined. In addition, keywords and classes of each source layer and view were reported in another shared spreadsheet. Tracking-spreadsheets were shared between other colleagues via the GoogleDrive network in order to guarantee that no conflict-copy would be generated when working simultaneously on the same spreadsheet.

2.2. Assessment of interoperability between MapX and other platforms

The degree of interoperability of each source was determined in agreement with the Data Integrity Assessment Framework adopted at MapX (Table 1). The main parameters that influence the interoperability of a platform of data with MapX are:

1) the reliability of the data provided by the source (e.g. quality control of the data, documentation, authority of the source);
2) the technical accessibility of the dataset (e.g. the services provided by the data platform, the type of data that is provided by the source, the compliance with the OGC standards);
3) the openness of the dataset (e.g. the type of licence that is associated to the datasets provided by the source, its accessibility in terms of user rights);
4) the sustainability of the source (e.g. the capacity to guarantee up to date data with time).

In few words "an authoritative spatial data infrastructure with open source policies has a higher interoperability with MapX compared with an unreliable local source of private non-spatial data."

Specifically, the authoritative degree of a source is among the most important parameters that influences the reliability of a dataset. The possibility to stream the data from one platform to another is, however, the most important technical parameter that influences the interoperability between infrastructures. Web platforms with GeoNode technology allow other platforms to visualize their data directly from their server. This factor guarantees the streamed data to be up to date and the destination platform to be performant. Obviously, the quality of the data in the GeoNode depends on the degree of maintenance of the data provider and its attention to data documentation and opensource policies. Platforms that only allow data to be downloaded and published on thirds have a decreased degree of interoperability with other platforms because they require data to be re-published in another server. The problems related to updates frequency in the destination platforms may be smoothed by the use of scripts that automatize the download of the data from the source and their publication in the destination server. Clearly, the level of readiness of datasets and their coherency throughout the different updates determine the complexity of the data-publication process in the destination server. For example, increasingly low readiness levels is expressed by data provided in the wrong coordinate system (MapX requires it to be in WGS-84), data whose attribute table is not clean, complete or spell-mistake free, and especially, by non-machine-readable data. Lower interoperability is shared by institutions and data owners in general that do not give automatic access to data unless they are required specifically through a private request. The level of data accessibility is important in terms of interoperability between platforms because some platforms may apply restrictions on the use of data and, more commonly, on data re-distribution. Finally, the update frequency of the platform has an influence on the quality of the data as it may affect data-pertinence if that kind of data is supposed to change frequently.

RELIABILITY	1	Is the data documented?
	2	Does the data provider have the relevant mandate?
	3	Is the data subject to internal quality control?
	4	Is the data subject to external quality control?
TECHNICAL ACCESSIBILITY	5	Is the data machine-readable?
	6	Is the data available online?
	7	Is the data available in a non-fragmented state?
	8	Does the data comply with OGC standards?
OPENNESS	9	Is the data publicly available?
	10	Is the data accessible to the public without discrimination of users?
	11	Is the data available free of charge?
	12	Is there an open licence for the data?
SUSTAINABILITY	13	Does the data provider have the capacity to manage and update the layer?
	14	Does the data provider have a long-term mandate and financial means?
	15	Can the data layer be dynamically streamed and shared using an API?
	16	Does the data support measuring of SDG targets?

Table 1: Questions of the Data Integrity Assessment Framework that are relevant for the assessment of the level of interoperability between MapX and the data providers.

2.3. Stability and User-experience testing

Two types of testing were effectuated:

- **Stability test** → Is the application bug-free? This aspect was tested by trying to re-create all possible scenarios of usage of the platform.
- **User-experience test** → Does the application respond correctly to the needs of users? Does the application provide the expected functionalities when it is employed?

Reports were published on github.com and frequently led to successful improvements of the web application.

2.4. Production of didactic material

Teaching material was developed on two major subjects: 1) the functionalities of MapX v. 3 and 2) the Github procedure for reporting issues with the application.

Both documents were thought for people that do not have any experience with MapX and only minimum skills on GIS. Documents had a cut that will allow users to transmit the acquired knowledge to thirds. Detailed explanations and specific exercises on various tasks allow new users to fully exploit the potential of MapX and to contribute to the improvement of the platform. To complete the user-experience of the attendants, the googlesheets conceived for tracking and evaluating the published source-layers in MapX were shared with them.

Other documents were generated by colleagues to teach general concepts on GIS and SDIs, the update process of source layers, and the implementation of story maps in MapX. These documents constitute together the basis for sharing the knowhow of MapX to stakeholders.

3. Results

3.1. Consolidation of MapX data platform

The source layers that I published for Afghanistan sum up to 17. From these layers 21 views were created. Other layers were published for World (1), Somalia (5, resulting in 7 views), South-Sudan (3) and Haiti (1). In total, I published 27 layers which resulted in 33 views (Figure 4).

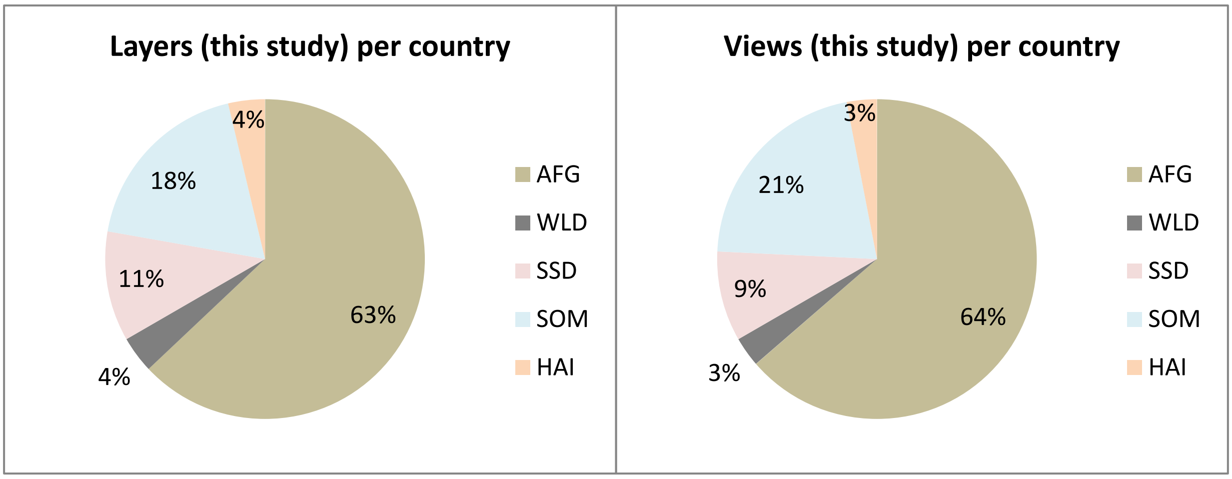


Figure 4: Percentage of layers and views published per country in this study.

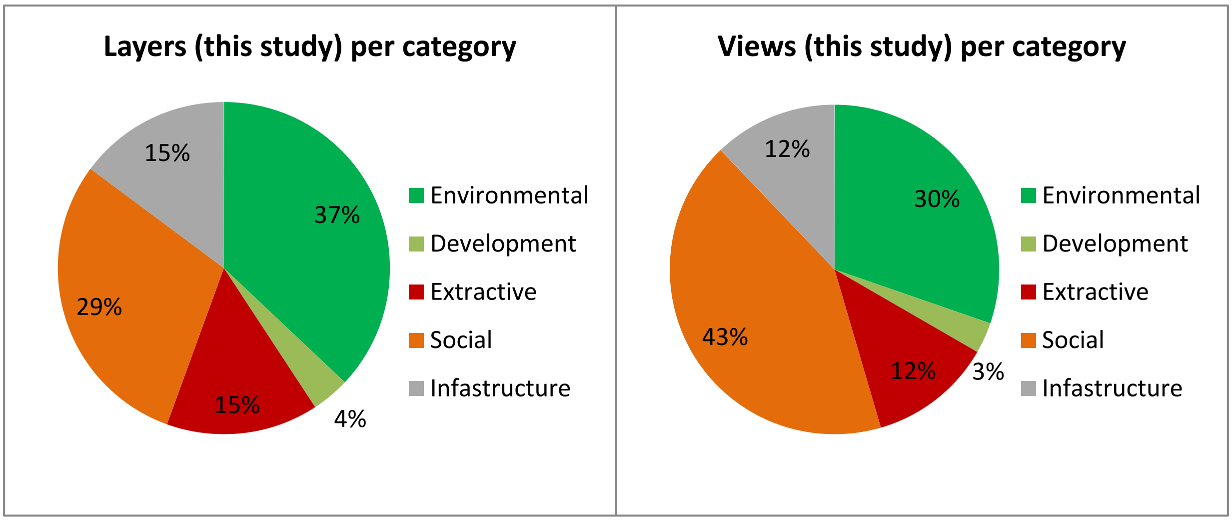


Figure 5: Distribution of published layers and views per category of data.

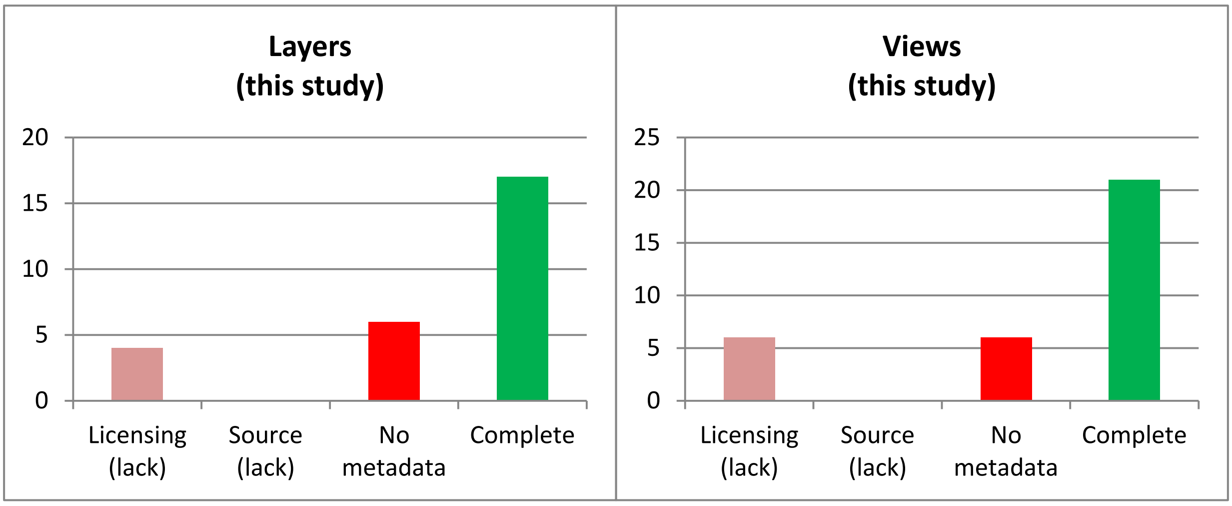


Figure 6: Metadata information per layer / view published in this study.

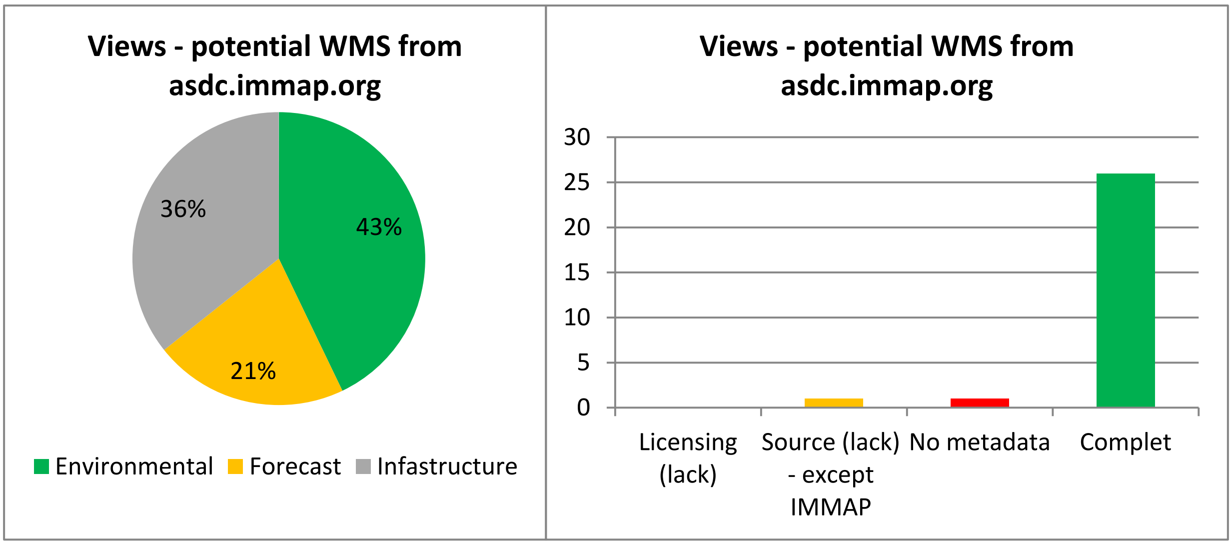


Figure 7: Afghanistan's risk related data selected from asdc.immap.org for streaming in MapX via WMS requests. Figures show the distribution of data per category and according to the status of their metadata.

Those source layers covered "environmental", "development", "extractive", "infrastructures" and "social" data-categories with largest amount of layers concerning the social domain (Figure 5). Due to the fact that multiple views were generated from single social layers, the number of views that concern social data further increased compared to the other domains with respect to the relative difference on source layers (Figure 5). Most of the layers published in this study have complete metadata while a small fraction misses licencing information and/or metadata entirely (Figure 6). This is similarly displayed by views (Figure 6).

Among the Afghanistan layers published during this internship, 4 layers (resulting in 4 views) were risk related datasets and concern the ongoing Disaster Risk Reduction (DDR) activities in Afghanistan. Besides this small fraction of datasets, a large amount of relevant data (28 views) that concern risk could be selected to be streamed in MapX from the spatial data infrastructure of iMMAP (http://asdc.immap.org/geoserver). At the moment, the security problems related to the absence of a certificate in the server of iMMAP did not allow the data to be available in MapX. As soon as iMMAP will solve this issue, the views in Figure 7 will be streamed into MapX and will be set available for public. Thanks to the high level of information provided by iMMAP, only a small fraction of the concerned views won't have their metadata completed (Figure 7). The level of consolidation of the platform for the Afghanistan project is shown in Figure 8 and Figure 9 and discussed later on.

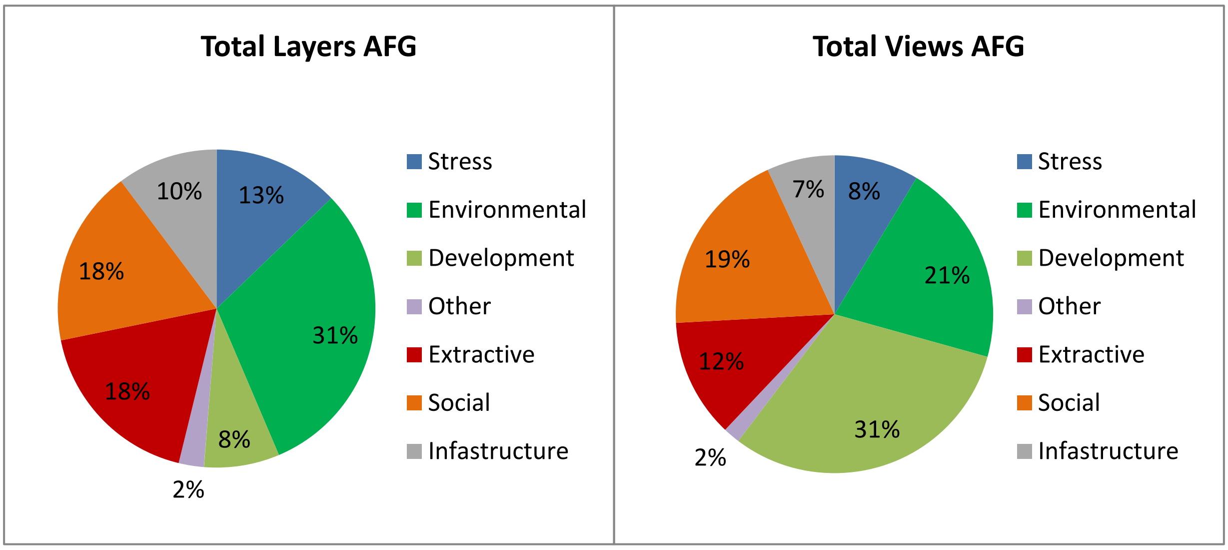


Figure 8: Published layers and views per category of data available for Afghanistan at October 2017.

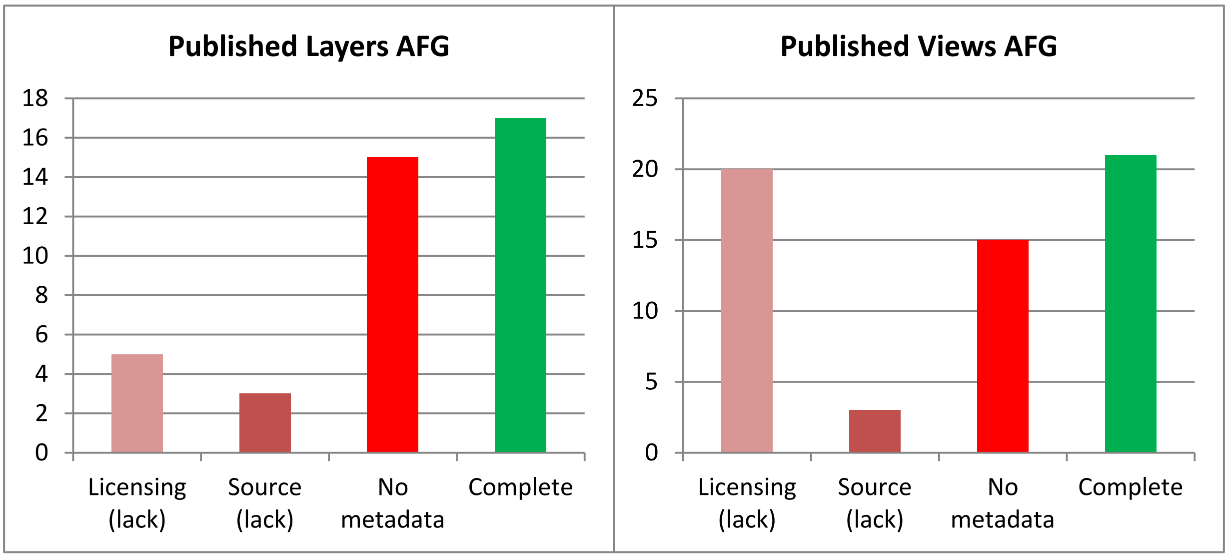


Figure 9: Metadata status for all available layers and views of Afghanistan at October 2017.

Automatization of the "date" to "POSIX" time-values conversion

```
# -*- coding: UTF-8 -*-
# Importing des modules systeme. If arcpy is not imported there is no access to arctoolbox
import arcpy, sys, os, string, datetime, time

# Lancement des toolbox requises
arcpy.AddToolbox ("C:/Program Files (x86)/ArcGIS/Desktop10.3/ArcToolbox/Toolboxes/Data Management Tools.tbx")
#This is to overwrite existing files when new ones are created
arcpy.overwriteoutput = True

#Workspace definition
WS = arcpy.env.workspace = "C:\\Users\\benve\\Documents\\UNEP Data\\04 UN ENVIRONMENT\\time_conversion_directory"

#Looking for available files to be edited
files = os.walk(WS).next()[2]
for file in files:
    if file.endswith(".shp"):
        file_name = os.path.splitext(file)[0]
        shape_in = WS + "\\" + file
        print (shape_in)
        feat_out = WS + "\\" + file_name + "_TimeConv"
        print (feat_out)

        #Making a feature class of the shp to work with
        arcpy.CopyFeatures_management(shape_in, feat_out)
        editing = feat_out + ".shp"
        #Detecting "date" attributes in the attributes' table of each feature class
        fieldslist = arcpy.ListFields(editing)
        i = int
        i=0
        for value in fieldslist:
            if value.type == "Date":
                date_fields = list()
                date_fields.append(value.name)
                var_name = "mx_t" + str(i)
                date_fields.append(var_name)
                arcpy.AddField_management(editing, var_name, "LONG")
                i=i+1
                with arcpy.da.UpdateCursor(editing, date_fields) as cursor:
                    for row in cursor:
                        row [1] = time.mktime(row[0].timetuple())
                        cursor.updateRow(row)
```

Figure 10: Python script for time-values conversion to POSIX format. In yellow, the actual line for the conversion. The script allows to automatically detecting shapefiles with "date"-type attributes, performs the time-conversion process, paste values in new "double" attributes in a copy of the original shapefile.

The MapX application offers the possibility to filter the data that is shown on map on the basis of selected attributes from the attributes table of each spatial data. Accordingly, attributes with time-related data can be used to filter the visualized data by time through a time-slider tool. The time-slider only works on attributes which values are expressed in the POSIX (Portable Operating System Interface) format.

POSIX is a family of standards specified by the IEEE Computer Society for maintaining compatibility between operating systems (The Open Group). It expresses time as the number of seconds that have elapsed since 00:00:00 Coordinated Universal Time (UTC), Thursday, 1 January 1970, minus the number of leap seconds that have taken place since then (Open Group Base Specification Issue 7).

The conversion of values of date attributes can be done on different environments such as R and Python in few code lines (Basic date and time types - Python). During this study, I have developed a Python script that can be used for converting date values to the POSIX format (Figure 10). The script allows to automatically detecting shapefiles that have at least an attribute in "date" format and generates a new shapefile with new attributes where the time values are converted in the POSIX format. The script automates the process of shapefile detection, attributes seeking, new attributes creation, and data conversion of all the "date"-type attributes. The new attributes are named mx_t0, mx_t1, etc… in an incremental fashion.

This script works with every shapefile independently from the number of attributes in "date" format. However, MapX only accepts files that are featured with 1 or 2 date-attributes which might identify date of the event or start date and end date of events. In case more than 2 attributes with date values are present, the generated shapefile will be unusable in MapX. Therefore, the unsuitable shapefiles must not be placed in the specified directory for time-conversion. Alternative ideas concerning this issue are proposed in 4.3.

3.2. Interoperability between MapX and other platforms

The Afghanistan project (detailed in Annexe 3)

Data received from UN Environment Afghanistan were in large number useful and worth to be published in MapX. However, a few problematics were encountered during the preparation and publication of the selected datasets for this study. Often, metadata information was incomplete. A few datasets required their coordinate system to be defined before being converted to WGS-84 via GIS software. More demanding editing was required by the datasets which attribute table contained spell-mistakes, useless attributes with no values and incomplete information, and by datasets that had to be merged with other datasets (e.g. road networks of Afghanistan). The unavailability of government websites on which these data might be stored and set available for download complicated the understanding of relevance of the datasets.

The highest levels of interoperability were assessed for the spatial data infrastructures of iMMAP (The Afghanistan Spatial Data Center). IMMAP offers a high quality service and can guarantee accessible, up-to-date and clean data to be streamed into MapX. However, only a reduced amount of spatial data is present in the GeoNode of iMMAP and cannot substitute the mole of data received by UN Environment Afghanistan.

Other sources of data that were used for the Afghanistan project were characterized by a lower degree of interoperability and did not only concern Afghanistan data. Among them, sources like WWF, FAO, WDPA and OCHA provided accessible, clean and ready datasets from their websites that could directly be published into MapX after being downloaded. Some sources provided useful datasets in non-spatial format (.csv or .xls) which required a jointure to a spatial dataset in order to visualize them. Specifically, this problem involved the datasets on estimated population. This was rendered into MapX by joining 1) the .csv data provided by the Office for the Coordination of Humanitarian Affairs (OCHA), but originated by the Central Statistics Organization (CSO) of Afghanistan, with 2) the spatial distribution of districts and provinces defined by the Afghan Geodesy and Cartography Head Office (AGCHO). This procedure was preferred to simply using the spatial dataset on population estimated received from UN Environment Afghanistan that was 1) out of date (estimates for 2002-2003) and 2) displayed through unofficial administrative subdivisions of districts and provinces.

Other relevant sources that were investigated during this study include USGS. The geological society of the United States offers absolutely high quality data with open-source licence for a large range of domains around Geology. A selection of the data that was received by USGS was also received by UN Environment Afghanistan but the integrity and update level of the latter were generally lower. Besides the necessity to demand the USGS data via private request to their email contact, the only real negative point of USGS is the low degree of organization of data in the provided directories. Furthermore, metadata are sometimes not accurate as they are employed for more than one dataset with a consequent low level of detail and a large amount of redundant information. These conditions made difficult choosing the right and most up-to-date dataset and its corresponding metadata for publishing in MapX. However, the large amount of available datasets, their high level of accuracy and up-to-date degree, combined with the overall high accuracy and detail of their metadata compared to those received from UN Environment Afghanistan, makes USGS a better source of information for MapX in terms of interoperability for geological data. Furthermore, USGS provides the relational databases together with the spatial data that were originated from. Therefore, other potential spatial datasets could be generated from the data provided by USGS.

The EITI financial data for the Democratic Republic of Congo (feasibility assessment, detailed in Annexe 4)

The EITI provided pdf documents per each year since 2007 where the financial data of a number of mining and hydrocarbon companies are reported. This research focused on the mining sector and aimed to investigate on the coherency of the data provided between the years 2007 and 2012. Due to the absence of reports for the years 2008 and 2009 (EITI-DRC), the study could only take into consideration the year 2007 and those between 2010 and 2012 (Table 2). Results of the assessment are summarized into the detailed report Annexe 4 and the spreadsheet Annexe 5 where the relevant tables of each report and the list of companies' names are provided per each year.

Overall, coherency between reports is high enough for envisaging the creation of a spatial dataset on the financial data that concern the mining companies and publish it in MapX. In fact, almost the totality of the concerned companies can be traced across the different reports although their name and/or acronym change sometimes. The combined use of acronyms and names mostly revealed a successful method that could diminish the effects of data weaknesses.

Year	2007	2008	2009	2010	2011	2012
Nb. companies	21	-	-	49	118 (35)	93

Table 2: Number of companies in the perimeter of the assessment of each year. When mentioned, the number of companies that are only featured with unilateral declarations (from government) is given bracketed.

The EITI dataset has been conceived with the following structure (Table 3): the selected parameters that should be included in the EITI dataset are the company's name, their Personal Tax ID (N.I.F.), the financial attributes, the year concerned and the activity of the company (production vs exploration). The financial attributes consist of the declaration of payments by companies (A), the declaration of receipts by State (B) and the absolute discrepancy between the two parameters (A-B).

Companies names	Companies Personal Tax ID (N.I.F.)	Declaration of payments by companies (A)	Declaration of receipts by State (B)	Absolute discrepancy (A-B)	Year	Stage (PR=Production, EX=Exploration)
XXX	NIF 1				2007	
.	.				.	
.	.				.	
.	.				.	
.	NIF 5				2007	
.	NIF 1				2008	
.	.				.	
.	.				.	
.	.				.	
.	NIF 10				2008	
.	NIF 1				2009	
.	.				.	
.	.				.	
.	.				.	
.	NIF 12				2009	
.	NIF 1				2010	
.	.				.	
.	.				.	
.	.				.	
.	NIF 25				2010	
.	NIF 1				2011	
.	.				.	
.	.				.	
.	.				.	
.	NIF 30				2011	
.	NIF 1				2012	
.	.				.	
.	.				.	
.	.				.	
.	NIF 50				2012	

Table 3: Plan of the attribute table conceived for the spatial dataset on EITI data for DRC.

According to the developer of the MapX application (Mr. Frederic Moser), the best way to include the EITI data into MapX would be in the form of the attribute table presented in Table 3. Companies must appear in more than one row if they are present in more than one report (more years) in order to guarantee all data to be in the same file. The repetition of companies in different rows of a spatial dataset will inevitably bring to the repetition of geometrical information which will result in heavy spatial datasets. In order to avoid this problem, the EITI table of data would be deprived of geographic information which will instead be published in the platform as another source layer consisting of point data. The choice of point data instead of polygons is fundamental in order to associate a value that is representative of the essence of each company and not of their areal extent. Furthermore, this choice helps providing one and one only value for each company, even for companies that are featured with more than one mining concession which result in being constituted of several polygons. The jointure between the text and the spatial datasets will be performed on demand when data will be visualized on map through a common attribute (e.g. the N.I.F.; Figure 11).

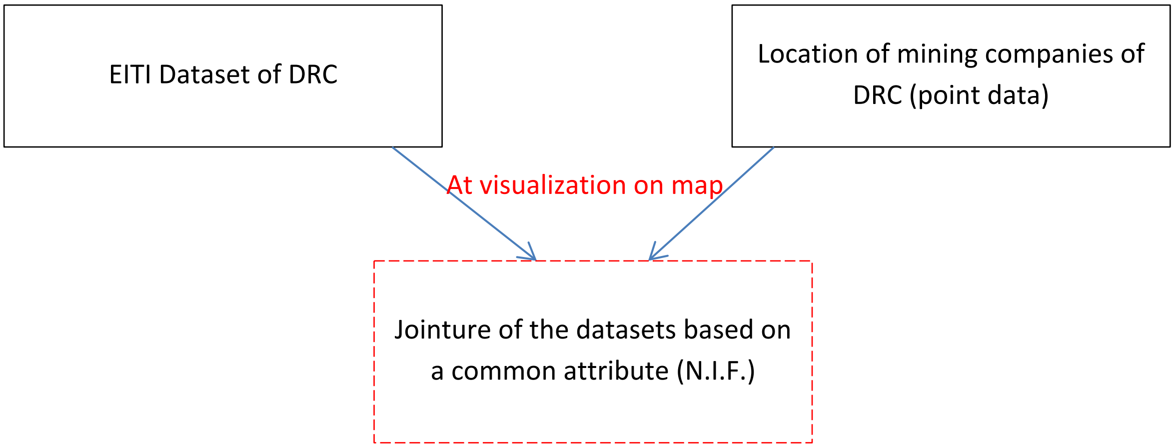


Figure 11: Schematic workflow of an ideal interaction of datasets in MapX for the EITI financial data of the DRC.

3.3. Stability and User-experience testing

The time dedicated to this activity helped the MapX web interface to be considerably improved by its developer.

In the time spent at GRID, I reported 29 times, 9 of which concern issues that were later fixed. In particular, reports concerned 9 times the platform stability and 20 times issues related to the user-experience (Figure 12). Among the latter, potential improvements of the platform were proposed 9 times through “feature requests”.

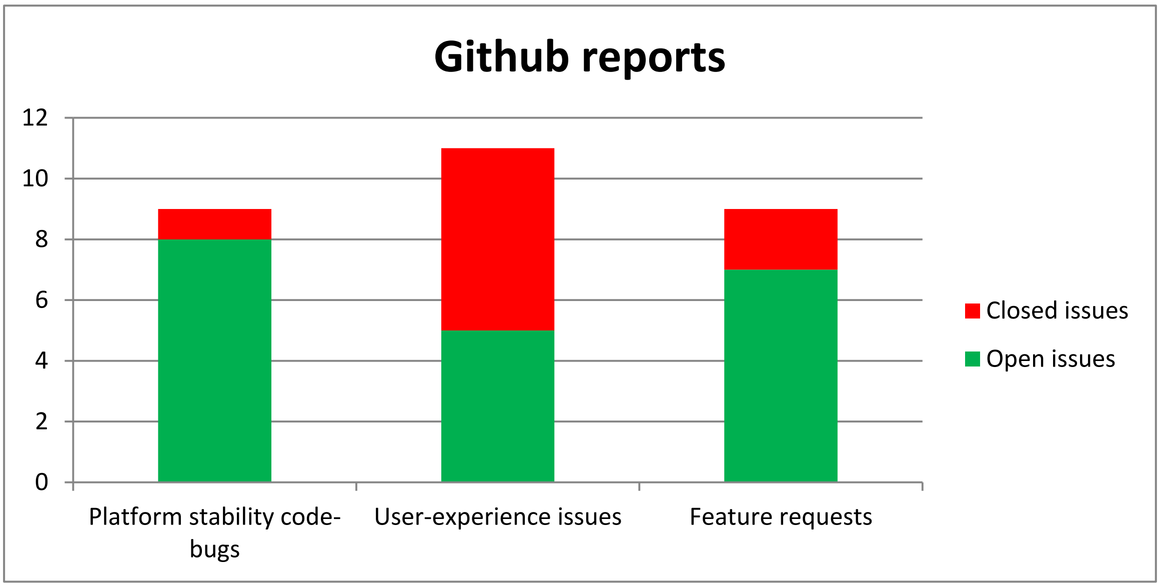


Figure 12: Github reports for MapX related issues and feature requests.

3.4. Production of teaching material

The conceived documents were redacted as text files in MS World. The document on MapX functionalities was conceived in different chapters from the most general use of the application to detailed explication on how to generate and configure views. A large number of screenshots were added to help users going through the different chapters. The document was repeatedly updated following the multiple updates of the interface of the platform and its functionalities.

The documents were distributed to the afghan partners in order to improve consolidation of the platform and stability-testing by enlarging the number of people using MapX. In particular, I used the document on MapX Key functionalities for direct teaching to the afghan partners at UN Environment – GRID in Geneva.

4. Discussion

4.1. Consolidation of MapX data platform

As a consequence of both the combined work of different colleagues at UN Environment - GRID in Geneva and the collaboration with UN Environment Afghanistan, Afghanistan is at the moment the country with the largest number of published source layers (39) and views (58). The list of published layers and those selected for WMS for Afghanistan is detailed in Annexe 6.

Among the published layers, most of them are environmental data; the others display stress, extractive, social and infrastructure related data while a small portion of data concerns locations of study areas (labelled as “Others”; Figure 8). Due to the large number of views that were created out of few development indicators, the number of views that concern social data overwhelms considerably the other views (Figure 8). Overall, most demanded data for Afghanistan were published into MapX. The most important missing datum concerns the mining concession areas for Afghanistan. Mining concession data could not be obtained by UN Environment Afghanistan despite their repeated contacts of the UN offices with the afghan government.

Among the afghan layers and views, a large part of them still lack metadata information (Figure 9). The absence of information to be associated to the published data reduces the authoritative power of MapX as this project cannot assess whether the information set available for public in the platform comes from an authoritative source or not.

In general, the MapX platform is definitely and rapidly increasing the number of published data to be managed. This suggests that large amounts of data will have to be managed after being uploaded the first time. Clearly, a system for keeping track of the updated source layers that stands on the MapX platform rather than on the googlesheets, which might be featured by incomplete information and errors, must be set up. Further improvements on MapX interoperability with other platforms are suggested further on in 4.3.

4.2. Interoperability assessment

The Afghanistan project

The data received by UN Environment Afghanistan were largely obtained by government sources whose websites are mostly not working at the moment except for the iMMAP website. This complicated the process of selection of the most relevant data when a number of similar datasets were received. Consequently, the interoperability with the platform of data of UN Environment Afghanistan was not excellent. The information on the datasets that were received relied exclusively on the metadata associated to them, which were in cases weak or absent. Furthermore, the version of the data that was received could have been a past version while a more up to date version of the dataset or simply a cleaner version of the attribute table would be available. This problem could not be overwhelmed by visiting the webpages of the data sources because the government websites associated to the sources of data were often unavailable (e.g. AIMS, Afghanistan Environmental Data Centre) or seemed to not giving access to the spatial data. In the actual conditions, it was difficult determining the quality and relevance of the available datasets. Fortunately, an evaluation of the data kind on GIS software often helped assessing which dataset had to be chosen for publication into MapX.

The interoperability between MapX and UN Environment Afghanistan governmental data could be definitely improved by re-activating the government websites and by providing public access to clean and up-to-date datasets and metadata on dedicated websites. A further improvement of the interoperability between government sources and MapX relies on the possibility to have data streamed into MapX through WMS requests such as with the iMMAP GeoNode. The iMMAP infrastructure is definitely characterized by the highest level of interoperability with MapX as it provides up-to-date and pertinent data that can be easily streamed into MapX without being required to update the datasets in the future in MapX. Nevertheless, the iMMAP website does not allow downloading the datasets (in case needed) although the download button is present. Data can anyway be downloaded from their GeoNode from asdc.immap.org/geoserver with an account.

Other web sources contributed to the process of platform consolidation with good quality and coherent data that could largely be directly published into MapX. The only potential improvement on the interoperability of these platforms would be related to the possibility to have data published automatically in MapX through scripts.

The EITI assessment for the DRC

For what concerns the EITI-DRC datasets, the level of interoperability of the data provider is not at decent levels and places itself at lower position compared to all other investigated sources. The EITI data require to be published into a spatial dataset and, furthermore, they need to be assembled from different year-specific files that are produced in an un-friendly format (.pdf). Fortunately, the level of coherence between the reports is high enough to allow companies' names to be tracked across the available reports.

The quality and time needed to prepare the data will strongly depend on the accuracy of the parameter needed for the unique identification of each company (which is proposed to be the N.I.F.). As this parameter is not given in the report for the year 2007, it is suggested to look for the corresponding companies' N.I.F.s in recent reports. This might work as companies that are present in the oldest reports are also present in the most recent ones. However, few spotted cases of companies being only mentioned in old reports may raise complications (Annexe 5). Furthermore, technical hitches are also expected to occur because of potential N.I.F. duplications (Arango 2017). This problem is related to the structure of large companies which have subsidiary companies under a different name that keep the N.I.F. of the parent company. A possible solution for this problem would be using another identifier for the mining companies or integrating the N.I.F. with an additional code that is only related to the companies' names. However, this might extend the time needed for processing the data. In general, if the process has to be performed one time only, the best solution is to manual clean all the available data (including the spatial data) in order to associate a unique ID (e.g. N.I.F._ACRONIM or ACRONIM if N.I.F. is not available), to each company. It is suggested that this process would be potentially satisfying due to the high level of coherency between reports that allows companies' names to be tracked across the available reports. Once the association is done, the process is straightforward. If the publication of these data is planned to be performed automatically every year, it may requires the manual intervention of someone. It is necessary to keep in mind that keeping a large dataset of the names associated to each country in all reports will help in future years the integration of new data into the database published previously in MapX.

4.3. Future perspectives on user-experience-based interoperability

During the three and a half months of this research MapX improved its interoperability with external sources of data through a large number of modifications. First of all, it allowed multipart geometries to be published (MultiPoint, MultiLineString & MultiPolygon). Furthermore, it is now possible to publish datasets that contain null geometries. These corrections allowed some datasets among the ones received from UN Environment Afghanistan to be included into the platform. Still, MapX only partially supports GeometryCollection datasets (e.g. lines together with points or polygons in the same dataset). Another important improvement is related to the possibility to upload large files which, however, still requires some corrections. Finally, concerning the WMS requests, it is now possible to have the legend of the view generated through a WMS request in the form of a png image.

In the close future, MapX will become the official platform for spatial data visualization of UN Environment – GRID on the domain of the natural resources. Therefore, large changes are expected to take place in its structure in order to guarantee services to users that habitually employ the platforms that will be replaced by MapX. Although this is a huge possibility for MapX to grow thanks to knowledge transferred from the existing platforms, it is envisaged that the constraints that the future changes will generate on MapX will not reduce the margins of evolution of the platform.

Said so, thanks to experience maturated in this internship, the following improvements are suggested for a continuous improvement of the user experience with the MapX web application and its interoperability with other platforms:

- Keep track of the source layers stored and streamed into MapX – and the views associated to them - via spreadsheets generated by MapX that report for each source layer, its country of publication, its category, the associated views and the countries where views are available;
- Allow spatial data that are available for download on web platforms to be automatically published into MapX, and views created out of them, through scripts. Allow the update of the datasets to be performed automatically.

- For datasets that need manual editing, allow the system to remind superusers which dataset is potentially out of date. To do so, the system could base its action on the field "Periodicity" among the metadata of the source layers;
- Allow metadata to be streamed together with data from external servers through WFS requests;
- Allow the ISO metadata to be integrated into the MapX json structure of metadata in order to decrease time for publishing source layers when .xml files are available;
- Allow vector data to be streamed as vector data in order to interact on map with the attribute table;
- Perform "date" to "posix" transformations into MapX in order to facilitate the streaming of datasets with time-related attributes;
- Allow choosing which "date"-type attribute/s must be used for time-slicing the data when configuring the view. This will allow datasets that are featured with more than 2 time-attributes to be converted in an automated way by the Python script at 3.1 and still be accepted by MapX. Furthermore, this process will exploit all the potential for time-filtering provided by datasets with more than two time-related attributes. This opportunity will facilitate the process of publication and employment of datasets that involve time-related attributes.
- Allow the "sld" and "lyr" files for style definition to be imported into MapX and to be used for views created out of the specific layers;
- Consider 2 variables for manual styling of point data in order to use both size and colour properties of points;
- Provide the style configuration window with colour scales to be selected as in ArcGIS in order to save time and guarantee styling to follow specific rules;
- Allow source layers with non-WGS-84 coordinate systems to be published into MapX and perform a coordinate transformation into MapX using GDAL libraries;
- Allow a view to be edited in whatever country it is shown if possible;
- In alternative, mention in which country is the view editable when a view is visualized;
- Allow the temporal and local views that are generated through the drag and drop of source layers to be configured and their style to be determined. People with no rights for publication will be able to visualize their datasets with their preferred style. This might open new domains to MapX.

Conclusions

In conclusion, this study allowed the platform of data MapX to be considerably consolidated with the publication of a large number of source layers and the genesis of an even more conspicuous number of views. The Afghanistan project constitutes the largest part of this study. This project provided a strong evidence of the value of the MapX application for data filtering and cross-visualization on the extractive-related domain. However, this project revealed also some negative points which must be improved if MapX wants to become the reference point of governments and different stakeholders in the domain of natural resources. Among these, the most important factor is related to the need of updating source layers when they become out of date. The exponentially rising number of countries joining MapX in the future, and the consequent increasing mole of data being uploaded into the platform, stress the necessity of updating datasets in MapX automatically. As most data providers do not have implemented a spatial data infrastructure to communicate with, it is suggested to concentrate on updating the source layers and views by downloading the spatial datasets from the websites and publishing them at the place of the old ones.

During this internship, several improvements of the platforms were proposed. Very interesting discussions were generated with Frederic Moser, Thomas Piller and Alejandra Arango. A few suggestions were implemented while others will potentially be in the future. In general, this part of the project considerably improved my knowledge on the management of a web application for spatial data visualization. The dynamism and interaction within the team was a strong stimulation that contributed to the continuous growing and evolution of the platform. I consider this a very positive experience and the domain of SDIs something to explore deeper.

References

ARANGO, ALEJANDRA, *Identification and preparation of geospatial data for their integration in the MAP-X online platform for the extractive sector*, research thesis, Complementary certificate on Geomatics, Geneva, 2017.

BASIC DATE AND TIME TYPES – PYTHON, https://docs.python.org/3/library/datetime.html.

EITI, https://eiti.org/.

EITI-DRC, www.itierdc.net.

GITHUB REPORTING PLATFORM, https://github.com/fxi/map-x-mgl/issues.

IMMAP, http://asdc.immap.org.

MAPX, https://app.mapx.org/.

MINAMATA PROJECT, http://mercuryconvention.org/.

OPEN GROUP BASE SPECIFICATION ISSUE 7, http://pubs.opengroup.org/onlinepubs/9699919799/basedefs/V1_chap04.html#tag_04_16.

THE OPEN GROUP, www.opengroup.org/austin/papers/posix_faq.html.

UN ENVIRONMENT, www.unep.org.

UN ENVIRONMENT-GRID, http://www.grid.unep.ch/index.php?option=com_content&view=article&id=1&Itemid=511&lang=f.