Deliverable 1.1

Data Management Plan

30th April 2016

Version 1.4

Abstract:
First release of the its4land Data Management Plan

Project Number: 687828
Work Package: 1
Lead: University of Twente
Type: Report
Dissemination: Public
Delivery Date: 01 May 2016
Actual Delivery Date: 30 April 2016
Contributors: Markus Gerke, Carl Schultz, Kaspar Kundert

Its 4 Land
Hengelosestraat 99
Enschede 7500AE
Netherlands
Phone: +31534874532
www.its4land.com

The its4land consortium consists of the following partners:

University of Twente (UT)
KU Leuven (KUL)
Westfaelische Wilhelms-Universitaet Muenster (WWU)
Hansa Luftbild AG (HL)
Institut d'Enseignement Superieur de Ruhengeri (INES)
Bahir Dar University (BDU)
Technical University of Kenya (TUK)
esri Rwanda (ESRI).

Summary

The its4land project participates in the “Open Research Data Pilot” and therefore maintains a Data Management Plan (DMP). This report is the initial DMP. It is in accordance to the guideline provided by the EC. It must be noted that this DMP is not considered final, this applies especially for chapter 3, which reports about individual datasets. In subsequent deliverables (reports) this DMP will be updated in parallel.

Contents

1 Introduction

The H2020 project *its4land* is participating in the "Open Research Data Pilot". To this end, the initial creation and maintenance of a Data Management Plan (DMP) is obligatory. According to the Grant Agreement, the initial DMP (Deliverable 1.1) is due in month 3. The DMP will be updated regularly and sent to the EC as an independent deliverable.

This document presents the initial DMP. It has several parts which are in accordance with the DMPonline tool, as advised in the corresponding EU documents. Some sections created within this tool, however, are of a general nature, in particular the technical data handling. Therefore, those parts are described first (Section 2), while in Section 3 the individual datasets are described. In those descriptions reference is made to the overall part, or additional remarks are given, when appropriate.

1.1 Project description in brief

The aim of its4land is to develop an innovative suite of land tenure recording tools inspired by geo-information technologies, that responds to end-user needs and market opportunities in Sub-Saharan Africa (SSA), specifically reinforcing an existing strategic collaboration between EU and SSA.

1.1.1 Objectives

- to capture the specific needs, market opportunities, and readiness of end-users in the domain of land tenure information recording
- to co-design, adapt, integrate, demonstrate, and validate a land tenure recording suite based on small unmanned aerial vehicles (UAVs), smart sketchmaps, automated feature extraction, and geocloud services
- to develop and valorise a governance model that realizes the innovation process by aligning end users' conditions, technological opportunity, business models, and capacity building requirements

1.2 its4land Information Structure

Data information management within its4land rests on 3 pillars: a) EU-directed project management, b) data production and storage, and c) results dissemination. While a) refers to the communication and data exchange between the project management and the EC (eg. through Sygma), b) resembles the project internal data infrastructure. Access to data is restricted to project partners. Public dissemination refers to our actions and measures for representing the project to people other than project partners. Figure 1 illustrates these three pillars and their relationship with other aspects of the project. The DMP refers mainly to b): data production and storage, but also partly to c) results dissemination.

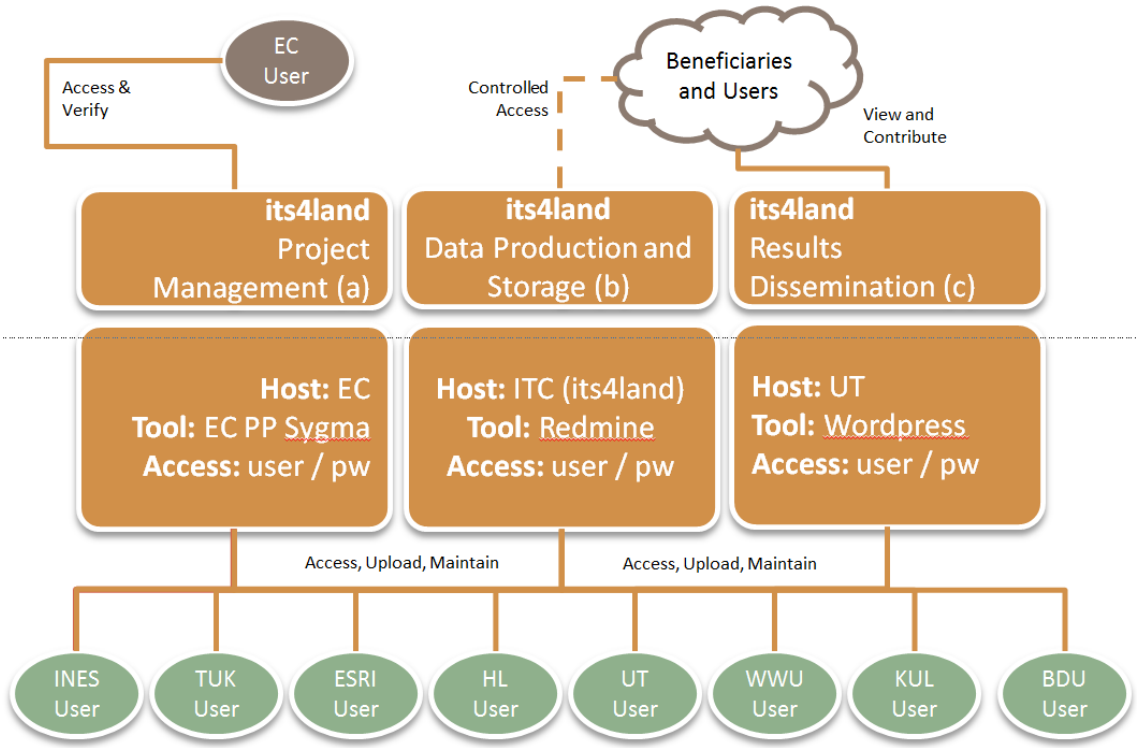

Figure 1: Data and Information pillars: project management (a), data production and storage (b), results dissemination (c)

2 General notes on data sharing, security, backup and archiving

The project manager and coordinator (ITC) are responsible for the technical data maintenance (pillars b and c), while pillar a (refer to Fig. 1) is handled by the Sygma-system. The latter is not discussed in this DMP.

2.1 Data production and storage (pillar b)

Within the consortium a dedicated server (network attached storage, NAS) is available to all partners. The NAS not only provides access to storage in the sense of a file server, but also comes with a data management tool that allows access and backup via several (secured) means. This server is located within the ITC building and is thus connected to the network of the University of Twente.

2.1.1 Data sharing

All datasets and documents are stored on the server and are accessible via a so-called project management tool (redmine[1]), which is published under an open source license. This tool is characterized through

- File storage: files (documents, datasets) can be organized in several hierarchies, including
  - ✓ Versioning
  - ✓ Meta-Data descriptions (format can be freely defined)

- User and permission schemes: users (i.e. the project participants) can be grouped and each individual group or user retrieves permissions according to usual levels (read only, read-write, administrator, etc). The user management will be done centrally at ITC
- Hyperlink access to individual files/directories: if needed, a URL can be shared with other users or externals to ease dissemination

Data and documents are organized generally per *working package* (WP), including individual files shared by participants, and deliverable documents in progress. Within the WP structure, a sub-division according to tasks might be appropriate.
Such a structure enables easy permission issuance, for instance partners involved in a certain WP will get full access rights, while others might only able to read the data. This is a common mechanism to prevent unintended data manipulation. Data security is enforced because the server frontend is only accessible via an encrypted https connection and only to the partners within the consortium.

2.1.2 Archiving and preservation (including storage and backup)

Several strategies are implemented to guarantee that no data gets lost. Since the mentioned project management software comes with a versioning system that is embedded in a mysql database, archiving of historic data takes place implicitly.

[1] http://www.redmine.org/

Therefore, it is only necessary to ensure that the entire database and the underlying files are kept safe. This is done in three ways.

- Use of a redundant storage system: the data server used runs a RAID system with redundancy enabled. This means that in case of hard drive failure, the system will still be operational and no data will be lost.
- Regular backup to an external hard drive: as the first security instance, the database and underlying data gets stored regularly (at least once a week) to external hard drives, which are located in the same room as the server
- Regular backup to a cloud service: in order to prevent data loss in case of emergency situations which make the physical data server inaccessible, data will be stored using a cloud service in regular intervals (once a week), as well. The final decision on which cloud service to choose is pending, although we tend towards the Hubic[2] service.

2.2 Results dissemination (pillar c)

The website (http://its4land.com) has been online since month 1 of the project. Besides general information on the project, partners, and so on, it contains a section "Dissemination" (https://its4land.com/dissemination-and-exploitation/). Within this section, scientific publications (open access) and selected datasets will be published, however, in most cases links to existing repositories will be sufficient. If the published item does not have a digital object identifier (DOI) already, e.g. assigned through a publisher, it will be created by the its4land consortium. By this means all released data is easier to link and can be put in different web-based catalogues.

2.3 Self-assessment questions

The Guidelines on Data Management in Horizon 2020[3] give some general information on how to setup the DMP. One major task is to describe the different datasets in detail, see next section (annex 1 of the document). In annex 2 of the guidelines document a couple of self-assessment questions are posed. In the following we respond to those questions. The purpose is to centrally summarize mean properties of our DMP.

Data produced in its4land is:

1. Discoverable

DMP question: Are the data and associated software produced and/or used in the project discoverable (and readily located), identifiable by means of a standard identification mechanism (e.g. Digital Object Identifier)?

[2] www.hubic.com

[3] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (last visited April 2016)

Its4land-answer: YES- publications and final datasets which are made available to the public are assigned a digital object identifier (DOI).

2. Accessible
DMP question: are the data and associated software produced and/or used in the project accessible and in what modalities, scope, licenses (e.g. licencing framework for research and education, embargo periods, commercial exploitation, etc.)?

Its4land-answer: Yes and partially - data produced will be made available to the public. In an earlier stage of the project on request, in a later stage (to be decided by the consortium) datasets will be made public as a whole. This has to take into account 1) ethical considerations (for example whether the personal privacy of participants or stakeholders are affected), and 2) any patent issues. Hence, each request will be handled within the Management Team (MT) of its4land – with reference to the relevant ethics documents (Work Package 9) and legal requirements (i.e. Consortium Agreement).

3. Assessable and intelligible
DMP question: Are the data and associated software produced and/or used in the project assessable for and intelligible to third parties in contexts such as scientific scrutiny and peer review (e.g. are the minimal datasets handled together with scientific papers for the purpose of peer review, are data is provided in a way that judgments can be made about their reliability and the competence of those who created them)?

Its4land-answer: Partially - restrictions as listed under requirement 2 apply here as well.

4. Useable beyond the original purpose for which it was collected
DMP question: Are the data and associated software produced and/or used in the project useable by third parties even long time after the collection of the data (e.g. is the data safely stored in certified repositories for long term preservation and curation; is it stored together with the minimum software, metadata and documentation to make it useful; is the data useful for the wider public needs and usable for the likely purposes of non-specialists)?

Its4land-answer: Partially - the aim is to keep the data server and the website online beyond the lifetime of the project. In this regard, interested researchers can access the public data, or request additional information from a contact person for the online repository. However, since there are no financial means available after the project ends, a minimum lifetime of the online repository cannot be guaranteed. Backups might then still be available from the contact person. Whether the data will be transferred to a trusted repository will be decided in a later project stage. Again, such availability is subject to ethical controls of the datasets (See Work Package 9).

5. Interoperable to specific quality standards
DMP question: Are the data and associated software produced and/or used in the project interoperable allowing data exchange between researchers, institutions, organisations, countries, etc. (e.g. adhering to standards for data annotation, data

exchange, compliant with available software applications, and allowing recombinations with different datasets from different origins)?

Its4land-answer: Yes, since the aim is to use standard file formats and interfaces for all data. For instance, images are stored in jpg or tif format, including standardized header information, like EXIF or GeoTIFF. Final text documents are provided in PDF format, or simple txt, geographic data will be provide in format, readable by many systems, e.g. as. KML or Shapefiles

3 Individual dataset descriptions

The section is composed according to the description in the Guidelines on Data Management in Horizon2020, see above. Each dataset is presented in a separate subsection that is structured according to the guide. We also provide information per subsection in the form of a checklist to assist readers in gaining a more comprehensive understanding of each dataset. The checklist is adapted from the DMP template provided by the University of Twente [4]. If applicable, reference to the general description in the previous section is provided.

3.1 UAV imagery data

3.1.1 Data set reference and name

Images captured with an unmanned aerial vehicle, including geo-referenced meta data

3.1.2 Data set description

o How will data be collected?
Unmanned aerial vehicles, equipped with a imagery sensor (i.e. camera)

o Will you also use pre-existing data? From where?
No

o What type of data will be collected? (measurements, observations, questionnaires, models, etc.)
images, including (approximate) georeferencing information, derived from a satellite navigation device mounted on the UAV.

o In what file formats?
Common image file formats like jpg or tif, ASCII text files for georeferencing information

o Which tools or software are needed to create, process and/or visualize the data?
Standard image viewers, state-of-the-art Photogrammetric image processing software

o Do the data have a specific character in terms of reproducibility, confidentiality (e.g. privacy, see next question), etc.? What does this mean for the management of the data?
Images show land cover

o What is the estimated total size of the data, and what growth rate? What is the estimated number of files and the maximum file size?
Per image flight around max 500 images, each image 10MB, so 5GB per flight, several dozens flights during the project. An exact number cannot be given at this stage.

[4] https://www.utwente.nl/ub/en/services/MAIN/research-datamanagement/rdm/datamanagement-plan/ (accessed in April 2016)

o How do you handle version control to maintain all changes that are made to the data?
After image capture no update needed/possible. If the same area is captured again, it is a new, independent dataset.

3.1.3 Standards and metadata
Common image file formats like jpg or tif, ASCII text files for georeferencing information are used. Metadata is also stored as standard EXIF or GeoTIFF header

3.1.4 Data sharing
Technically the common description above applies.

o How will you manage copyright and Intellectual Property Rights issues? E.g. Who owns the data? How will the data be licensed for reuse?
Decisions on IP need to be discussed within the consortium/MT

o Are there any limitations on the access of your data?
Yes, see next questions

o What are the access criteria for the data (open/restricted access, embargo period, etc.)?
During runtime of project only accessible to project members, decision for later situation needs to be taken

o Who controls data access (e.g. PI Principal Investigator, student, lab, university, funder)?
WP leader or person in charge, nominated by the WP leader

3.1.5 Archiving and preservation (including storage and backup)
The common description above applies.

3.2 Stakeholders needs data analysis

3.2.1 Data set reference and name
The Stakeholders need analysis data are captured via semi-structured interviews, focus group and (online) survey

3.2.2 Data set description

o How will data be collected?
Literature study, semi-structured interviews, focus group and (online) survey

o Will you also use pre-existing data? From where?
Yes, data from relevant previous studies (e.g. clarified during the process of the research design) will be used

o What type of data will be collected? (measurements, observations, questionnaires, models, etc.)
Existing data from previous studies, interview data and questionnaire

o In what file formats?
Mainly in MS Word formats

o Which tools or software are needed to create, process and/or visualize the data?
MS Word and NVIVO for textual coding

o Do the data have a specific character in terms of reproducibility, confidentiality (e.g. privacy, see next question), etc.? What does this mean for the management of the data?
Yes, Most of the data are privacy sensitive – so accessibility of the data has to be discussed at the MT – with reference to its4land ethical frameworks (See Work Package 9)

o What is the estimated total size of the data, and what growth rate? What is the estimated number of files and the maximum file size?
The estimated size of the data is not so high, but the growth can be significant during the time span of the project (WP2). The estimated number of files will be somewhere in the thousands

o How do you handle version control to maintain all changes that are made to the data?
No relevant for this type of data

3.2.3 Standards and metadata
MS Word is used. Metadata is also stored, its format will be defined.

3.2.4 Data sharing
Technically the common description above applies.

o How will you manage copyright and Intellectual Property Rights issues? E.g. Who owns the data? How will the data be licensed for reuse?
Decisions on IP need to be discussed within the consortium / MT

o Are there any limitations on the access of your data?
No – but privacy conditions need to be respected

o What are the access criteria for the data (open/restricted access, embargo period, etc.)?
During runtime of project only accessible to project members, decision for later situation needs to be taken

o Who controls data access (e.g. PI Principal Investigator, student, lab, university, funder)?
WP leader or person in charge, nominated by the WP leader

3.2.5 Archiving and preservation (including storage and backup)

The common description above applies.

3.3 Sketch maps and ontologies for land rights and tenure records

3.3.1 Data set reference and name

Vector and raster images of sketch maps drawn on either paper or handheld tablets. A corpus of symbols for a visual language. Written descriptions of terms/concepts used in identifying land parcels. Audio recordings of interviews.

3.3.2 Data set description

o How will data be collected?
Via interviews and exercises with groups and individual citizens in the study areas. Pen and paper, digital pens, tablets, and voice recorders will be used. Base maps will be derived from data from various sources.

o Will you also use pre-existing data? From where?
No

o What type of data will be collected? (measurements, observations, questionnaires, models, etc.)
Sketch maps and verbal and/or textual descriptions obtained through group or individual exercises. Questionnaires will also be used
o In what file formats?
Common raster image file formats like JPG or TIFF, vector image file formats SVG and shape files, XML and JSON files for structured data, ASCII text files for unstructured data, mp3 files for audio data.

o Which tools or software are needed to create, process and/or visualize the data?
Standard image viewers, custom file processors, shape and symbol recognition tools.

o Do the data have a specific character in terms of reproducibility, confidentiality (e.g. privacy, see next question), etc.? What does this mean for the management of the data?
Sketch maps and textual descriptions will describe individual rights and therefore privacy issues must be addressed in the data storage.

o What is the estimated total size of the data, and what growth rate? What is the estimated number of files and the maximum file size?
The data we will collect will vary widely. Per participant per sketching exercise we estimate a lower bound on the data size at about 20MB including scanned images (8 MB), audio files (10 MB), vector data (1 MB) and metadata (1 MB). With an estimated 100 - 120 participants, we will have only about 2.4 GB but the data will likely grow post processing by a factor of 10 at most for a total of 24 GB. Base maps may consume an additional 20+ GB giving 50 GB in total.

o How do you handle version control to maintain all changes that are made to the data?
For the base maps, change reports will be attached to each new map generated. A list of base maps prepared for each exercise and task will be stored in a table with data such as date, task, etc. No automated versioning will be used.

3.3.3 Standards and metadata

Common raster image file formats like JPG or TIFF, vector image file formats SVG and shape files, XML and JSON files for structured data, ASCII text files for unstructured data, mp3 files for audio data. Embedded in each file.

3.3.4 Data sharing

Technically the common description above applies.

o How will you manage copyright and Intellectual Property Rights issues? E.g. Who owns the data? How will the data be licensed for reuse?
Decisions on IP need to be discussed within the consortium / MT

o Are there any limitations on the access of your data?
No – but privacy considerations need to be taken into account

o What are the access criteria for the data (open/restricted access, embargo period, etc.)?
A selection of appropriately anonymised data may be open to access. The default is that data are only accessible to project members during the course of the project.

o Who controls data access (e.g. PI Principal Investigator, student, lab, university, funder)?
WP leader or person in charge, nominated by the WP leader

3.3.5 Archiving and preservation (including storage and backup)

The common description above applies.