ISRIC
World Soil Information

Review of possible information platforms for CIRCASA's Knowledge Information System

ISRIC Report 2018/02

Niels H. Batjes

Director, ISRIC - World Soil Information
Droevendaalsesteeg 3 (building 101)
6708 PB Wageningen
The Netherlands
E-mail: soils@isric.org

Citation

Batjes N.H., 2018. Review of possible information platforms for CIRCASA's Knowledge Information System. Report 2018/02, ISRIC – World Soil Information, Wageningen (doi: 10.17027/isric-7Y7B-6S67)

Review of possible information platforms for CIRCASA's Knowledge Information System

Niels H. Batjes

ISRIC Report 2018/02

Wageningen, October 2018

Contents

List of figures

Preface

Many studies have shown that agricultural soils have potential to sequester carbon, when judiciously managed, resulting in various initiatives to sequester organic carbon in agricultural lands. Yet, there is still much uncertainty about this potential. The EU H2020 CIRCASA project aims to address this knowledge gap.

The CIRCASA project is implemented by a wide range of partners and coordinated by INRA (Institut national de la recherche agronomique, France). ISRIC - World Soil Information is tasked with the development of a Knowledge Information System (KIS) that will host knowledge on carbon sequestration in agricultural soils. This KIS will include metadata and data from experiments, as well as models and methodological guidelines developed by CIRCASA. ISRIC will also take part in the development of an On-line Collaborative Platform (OCP), a networking tool aiming at bringing together researchers, stakeholders and practitioners in the field. Through these activities, ISRIC will support the improved exchange and accessibility of information on carbon sequestration in agricultural soils.

As a first deliverable, ISRIC prepared a technical report specifying key requirements for developing a KIS for CIRCASA; in this approach, a new comprehensive platform would essentially have to be developed. Pragmatically, however, there are already several operational platforms that could be used as a basis for the knowledge information system. This report aims to provide a brief review of such platforms to inform the decision process.

ir. Rik van Den Bosch
Director, ISRIC – World Soil Information

1 Introduction

Awareness that agricultural soils have potential to sequester carbon has resulted in various initiatives to sequester organic carbon in agricultural lands (Soussana *et al.* 2017; UNCCD 2017; UNEP 2012). Yet, there is ongoing dialogue on the potential for sequestration in agricultural lands and a need for sharing knowledge and experiences on how to make this happen (FAO and ITPS 2015; Harden *et al.* 2017; Sulman *et al.* 2018). The EU H2020 CIRCASA project (2017-2020) aims to address this knowledge gap.

The overall objective of CIRCASA is to strengthen synergies among researchers and promote the transfer of knowledge on carbon sequestration in agricultural soils. As indicated on the CIRCASA[1] website, this will be achieved through four complementary activities that may be summarised as follows: a) strengthen the international research community, b) improve our understanding, c) co-design a strategic agenda, and d) create an international research consortium. The project, coordinated by INRA, is implemented by 24 partners.

Through the above activities CIRCASA will contribute to the implementation of the 2030 Agenda for Sustainable Development and the Paris agreement within the UN Framework Convention on Climate Change (UNFCCC). CIRCASA will achieve these goals in synergy with several international initiatives, such as The Global Alliance on Agricultural Greenhouse Gasses (GRA), the Joint Programming Initiative on Sustainable Agriculture, Food Security and Climate Change (FACCE-JPI), and the four per 1000 Soils for Food Security and Climate Initiative (4p1000). Further, CIRCASA will benefit from the CGIAR research programmes on Climate Change Agriculture and Food Security (CCAFS) and Water, Lands and Ecosystems (WLE).

The CIRCASA project will better structure current knowledge on agricultural soil carbon by:

i) Knowledge synthesis activities combining data (e.g. meta-analysis of scientific literature), integrating modelling results, developing methodological guidelines (e.g. on monitoring, reporting and verifying changes in agricultural soil carbon stocks),
ii) Co-designing with stakeholders a pilot knowledge system on agricultural SOC sequestration integrated into the Online Collaborative Platform (OCP). This integrated system will include geo-referenced meta-data and (when possible) data from experiments, observations and surveys, as well as from models and synthesis activities and methodological guidelines developed by CIRCASA. Ultimately, these tasks will lead to an enhanced international knowledge system (KIS) delivering improved scientific resources of both global and local significance (e.g. maps showing the technological potential for SOC sequestration of diverse agricultural practices).

Ultimately, the goal is to enable seamless integration of the discovered data with models and analytical tools. Users should be able to find easily resources via the OCP and to explore them visually. A prototype KIS will be set up for this purpose; it will provide access to geo-referenced meta data describing data from experiments, observations, surveys, crowd-sourcing, as well as from models and knowledge synthesis.

[1] https://www.circasa-project.eu/

The requirements for the KIS are described in CIRCASA deliverable D1.2 (CIRCASA/ISRIC 2018); this is a solid yet rather technical and conceptual document. The elaborate system would be delivered by ISRIC in month 24 of the project. However, following on discussions with the CIRCASA secretariat, ISRIC was asked to undertake a short review of freely available platforms for hosting metadata that may later be queried from the OCP. By their nature, these platforms may be less comprehensive than specified in the above 'requirements report'.

This report consists of four sections. Main requirements for the KIS as inventoried during an earlier study (CIRCASA/ISRIC 2018), hereafter referred to as DocD1.2, are summarised in Section 2. Subsequently, in Section 3, four potential open-source systems (metadata catalogues) for storing metadata in an inter-operable way, are discussed, pointing at their respective merits and possible limitations in relation to the overall CIRCASA objective. The need for a consistent ontology, and possible options for this, is discussed in Section 4. Concluding remarks and follow up actions are formulated in Section 5.

2 Requirements

Requirements of records to be considered in the KIS are detailed in Appendix A of DocD1.2, as prepared by Luis de Sousa (ISRIC) based on consultations with the CIRCASA consortium. These include: name (description of the requirement), priority (degree of importance, following the MoSCow scale: Must Have, Should Have, Could Have, Won't Have this time); type; subject (user domain, user profiles, processes); life phase; risk level (risk of failing to meet the requirement); rationale (for the requirement, textual); verifiable (a Boolean indicating whether the requirement is verifiable or not); and verification (indicating how the requirement can be verified).

The system will focus primarily on scientific information to support the scientific process and the development, management and evaluation of SOC policy, while other knowledge assets, that may be easier to understand by the general public, will be shared among network members using the emerging information sharing tools of the OCP. By its nature, the KIS is *not aimed at storing data, but rather at referencing and facilitating on-line access to the various resources, as held in various platforms* (see Section 3), using automatic search/download through queryable metadata. DocD1.2 requires adoption of OGC (Open Geospatial Consortium) web service standards (WMS, WCS and WFS) in this regard.

Examples of 'collaborative platforms' in the agricultural/environmental domain include those being implemented for the '4p1000' initiative[2] and by the 15 CGIAR research centres to 'mobilize the vast amounts of agricultural research data at their centres and elsewhere to produce new insights and increase the impact of agricultural research for development.'[3]

Another indispensable element of a dataset is its License as it determines whether a dataset is publicly available or not, which sets the conditions for its use in scientific studies or policy development.

[2] https://hub.4p1000.org/
[3] https://foodtank-com.cdn.ampproject.org/c/s/foodtank.com/news/2018/08/cgiar-big-data-platform-medha-devare/amp/

The number and types of knowledge assets referred by the central KIS is likely to become large, requiring a defined ontology (see section 3.2 in DocD1.2). By its nature, such an ontology will be dynamic; expert users will be expected to expand the pool of keywords through a supervised process as described in DocD1.2 (3.2). However, as the development and testing of a full-fledged ontology is a huge task, CIRCASA may wish to consider adopting an operational multi-lingual system (see Section 4).

A key process for any KIS is the submission of new datasets or publications. Data providers must first check whether the material they wish to submit is already accessible on-line. In the affirmative, a reference to its location (e.g. DOI) or to a service providing access (i.e. URL, ftp address or WFS connection string), will suffice to submit the dataset. Alternatively, data providers must make sure their dataset complies with the rules set by the Technical Platform Committee (TCP, see DocD1.2). These concern characteristics aspects such as spatial resolution, coordinate reference system, and data size/volume. Thereafter, the user has to define the meta-data. A minimum set of meta-data fields is desired in the meta-data record, and these may vary with the standard adopted for various existing platforms, as discussed in Section 3. As indicated, the metadata should clearly specify the licence according to which the data/information is shared (i.e. restricted to open access) and cited.

As indicated, the KIS is one of various deliverables of CIRCASA and it must be properly aligned with the OCP, the online platform that will function as the doorway to the CIRCASA knowledge base. The KIS must be integrated with the OCP in a transparent and seamless way, with users largely unaware of the transition between the two, even unaware of the existence of different technical platforms.

3 Review of possible data platforms

Four freely accessible on-line platforms are reviewed in this section in terms of their suitability for CIRCASA. To guarantee the long-term persistence of the ultimate system, such platforms should be based on open source technologies and be themselves open source. Thereby, future development and maintenance of the system will remain possible once the CIRCASA project ends in 2020. Dedicated communities will then be able to maintain and further develop the platform thus ensuring that the knowledge collected during the project remains freely available to the international community.

Metadata for peer-reviewed publications can be harvested from various sources through their DOI's, based on selected keywords. These platforms, such as Web of Science, are not discussed here.

3.1 Dataverse

Dataverse[4] is an open source web application to share, preserve, cite, explore, and analyse research data. It facilitates making data available to others, and allows users to replicate others' work more easily. Researchers,

[4] https://dataverse.org/about

journals, data authors, publishers, data distributors, and affiliated institutions all receive credit and web visibility through a Dataverse.

A Dataverse repository is the software installation, which may then host multiple virtual archives called dataverses. Each dataverse contains datasets, and each dataset contains descriptive metadata and data files (including documentation and code that accompany the data). As an organizing method, dataverses may also contain other dataverses.

Dataverse has grown considerably over time and is now a major international collaborative project. At the moment, there are 34 installations worldwide, corresponding with some 2872 Dataverses holding over 52,000 datasets from various domains. Some of them are 'restrictive' in the sense that they only accommodate data from a given institution, or in collaboration with a given institution, as illustrated in Figure 1. A demo version is available for testing purposes[5].

According to Dataverse[6], 'any type of quantitative or qualitative data (in any format)' can be uploaded. However, in practice, geospatial data may only be submitted as ESRI® shapefiles, thus excluding a wide range of other geospatial data (polygon and grid based).

By default, Dataverse uses the Creative Commons CC0[7] waiver for all submitted datasets (4.0 and on). However, importantly, data depositors can opt-out of using the CC0 waiver for their datasets, if needed, making full use of the available CC options[8] (e.g. CC-BY, by attribution). Metatadata requirements, however, remain the same, irrespective of the licence.

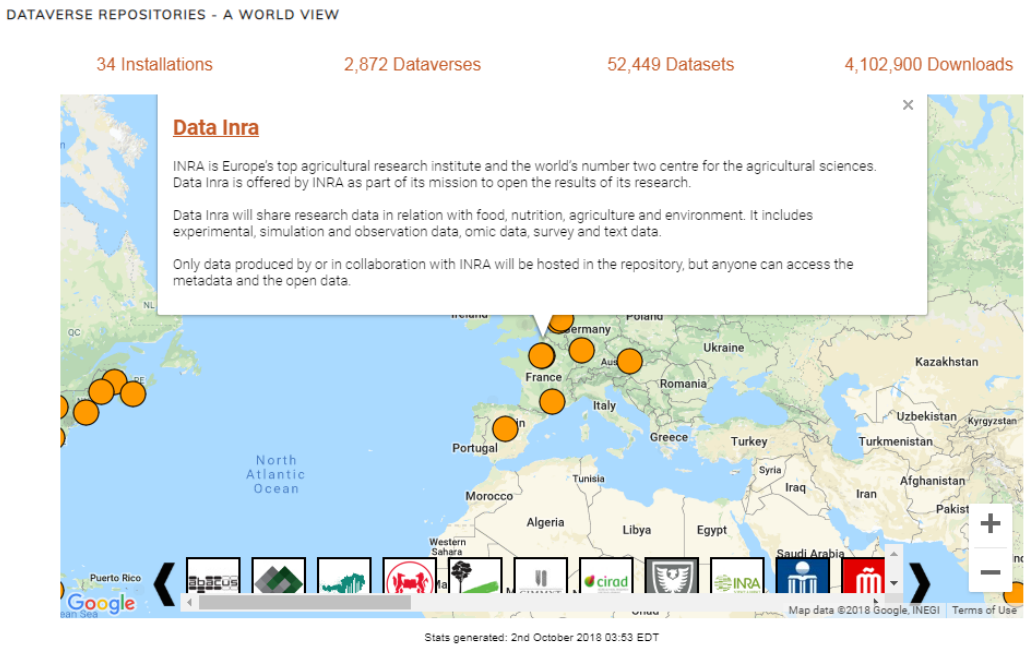


Figure 1. Example of a Dataverse (INRA, France).

[5] https://demo.dataverse.org/
[6] http://dataverse.harvard.edu
[7] https://creativecommons.org/share-your-work/public-domain/cc0/
[8] https://creativecommons.org/licenses/

Various integrations are available through which the functionality of Dataverse is gradually expanded, for example linkage to Matomo/PIWIK[9] and ORCID[10], see elsewhere[11]. Geospatial files may be mapped through integration with WorldMap[12]; as indicated, this for visualisation of shapefiles only. WorldMap, however, is itself a data harvesting system, sending all datasets to a server in Harvard; this could conflict with dataset licences.

A dataset in Dataverse is a container for data, documentation, code, and the metadata describing this dataset (Figure 2); main features are summarised in a screen dump (Figure 3). Information on functionality[13] and a developer's guide[14] are available online.

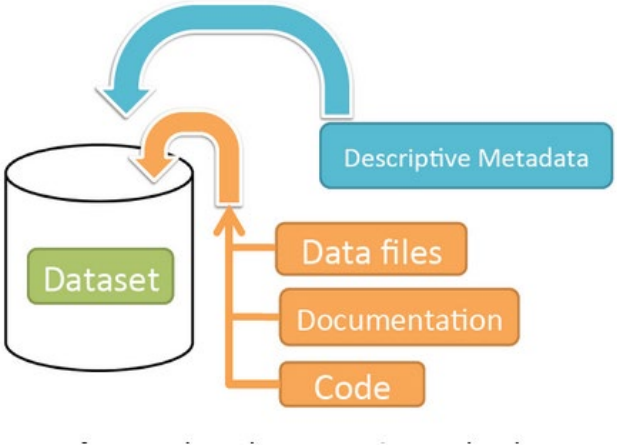


Figure 2. Key metadata fields of a Dataverse (https://dataverse.org/)

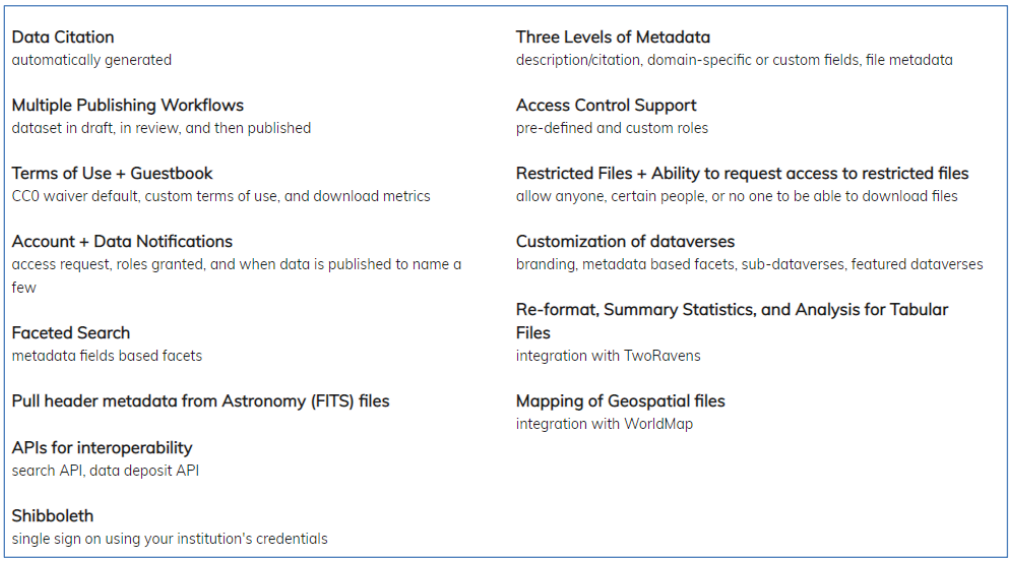


Figure 3. Main features of a Dataverse.

[9] https://en.wikipedia.org/wiki/Matomo_(software)
[10] https://orcid.org/
[11] https://dataverse.org/integrations
[12] http://worldmap.harvard.edu
[13] http://guides.dataverse.org/en/4.9.2/user/dataset-management.html#file-handling-uploading
[14] http://guides.dataverse.org/en/4.9.2/developers/index.html

Overall, Dataverse appears to be a well-supported and user-friendly system for managing a wide range of data, except geo-spatial data. As such, it is widely used worldwide. A possible shortcoming for CIRCASA, however, is that Dataverses cannot (yet) handle grid data, such as SoilGrids250m-derived SOC maps, but this may be remedied by using other (complementary) open source platforms such as GeoNetwork and Geonode. Further, Dataverse has its own Metadata format[15] whereas ISO or OGC standards may be preferred as indicated in DocD1.2.

If selected as a suitable candidate platform for CIRCASA, a project specific (customised) Dataverse may be considered for partners to add their data respectively to harvest project-relevant data from other Dataverses. Pragmatically, the central platform should be multilingual to facilitate international collaboration, rather than say mainly in French[16] as is the case for the INRA Dataverse; alternatively, metadata are provided using either French or English, which may be provide a source of confusion (see Section 4). Further, a drawback of Dataverses is that they are not (yet) OGC compliant, a key requirement listed in DocD1.2 (CIRCASA/ISRIC 2018).

Dataverses can be searched by broad subjects only, for example 'soils and soil sciences', 'water resources', 'farming systems and practices' or 'climate'. Further 'keyword terms' can be entered, but no specific thesaurus or ontology is mentioned. Such would be needed for more in-depth querying and data harvesting (i.e. inter-operability), as described by Madalli (2015).

An important feature of any data platform for CIRCASA is that the metadata it holds can be queried from the OCP. This involves as a process of exchanging metadata with other repositories. As a harvesting client, a given Dataverse can gather metadata records from remote sources. These can be other Dataverse instances or other archives that support OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting)[17], the standard harvesting protocol. Harvested metadata records will be indexed and made searchable by other users. Clicking on a harvested dataset in the search results will take the user to the original repository. Inherently, harvested datasets can only be edited in the source Dataverse installation, *in casu* by the original (authenticated) data providers. In cases, these may have to update their keywords to fit with the agreed upon ontology.

Special attention will need to be paid to user management in the selected KIS and its role in the integration with the OCP.

3.2 PANGEA

PANGAEA® is a data publisher, the services of which are generally open for archiving, publishing and re-usage of data. The World Data Center PANGAEA[18] is a regular member of the ICSU World Data System. As such, it is essentially different in scope from the distributed Dataverse system.

[15] http://guides.dataverse.org/en/latest/user/dataset-management.html#supported-metadata
[16] https://www6.inra.fr/datapartage/Gerer/Stocker-les-donnees/Portail-Data-Inra
[17] https://www.openarchives.org/pmh/
[18] https://www.pangaea.de/submit/

As indicated on the PANGEA website, any data from earth and life sciences are accepted for archiving. When starting the data submission process, data providers are redirected to the PANGAEA issue tracker that will assist them in providing metadata and uploading data files. Communication with PANGEA editors go through the issue tracker.

PANGAEA is archiving and publishing data according to the Open Access Model. Basic operation is covered through public funding, but PANGEA has to seek for additional funds, in particular for preparing and archiving new data. In case data are submitted through a project which has publication costs as part of the funding, PANGAEA would appreciate a small financial contribution for a data submission; other forms of funded collaborations can be negotiated.

Most of the data managed in PANGEA are freely available and can be used under the terms of the CC license mentioned on the data set description. A few password protected data sets are under moratorium from ongoing projects. The description of each data set is always visible and includes the principle investigator who can be contacted for access to the data.

PANGAEA is an archive for any kind of data from earth system research and thus has no special format requirements for submissions. Data may be submitted in the authors format and will be converted to the final import and publication format by the PANGAEA editors. Data providers are requested to keep the following points in mind to minimize the preparatory work prior to upload:

- For samples taken or measurements made somewhere on earth, *the provision of position(s) is mandatory* (latitude/longitude in decimal degree is preferred).
- If data are supplementary to a publication, the (preliminary) *citation with journal title and abstract* must be added.
- Submit ONE issue per publication supplement; several files can be attached to one issue (max. size per file = 100 MB)
- If data are related to a project (where PANGAEA is the designated archive) add the *project acronym* as label.
- Date/Time must be provided in ISO-format (e.g. 1954-04-07T13:34:11).
- Parameters are always accompanied by a *unit.*
- Abbreviations should be explained.
- Extended documentations may be added as plain text or pdf-file.
- Submit data tables as excel or tab-delimited text files; specific formats (e.g. shape, netCDF, segy ...) may be added in zip-archive. It is not apparent if large grid data sets can be processed.
- Submit via the PANGAEA ticket system.

As indicated, all data and metadata are quality checked, harmonized, and processed for machine readability by PANGEA editors, unlike for submissions to Dataverses.

PANGAEA offers a wide range of web services (SOAP / REST). This includes OAI-PMH for metadata harvesting, which is an important feature if metadata are to be accessed by the OCP. The API allows to retrieve any set of numerical and textual data. All PANGAEA datasets also ship with Schema.org / Dataset metadata.

It should be noted that PANGAEA supplies metadata to the data portal of the ICSU World Data System and takes care of the WDS search engine (ICSU-WDS).

Visualisation of georeferenced data in PANGEA through GIS (Geographical Information System) functionality is enabled by using Google Earth. Whether this functionality meets the specific requirements for CIRCASA's KIS (DocD1.2), remains to be assessed by an IT expert.

During data entry keywords can be assigned to a data submission. Apparently, there is no prescribed thesaurus for this in PANGEA. Further, 'keywords can be freely defined by the curator'[19].

Licences for all submissions are to be provided according to Creative Commons (version 4.0 International) except for CC0 (1.0 Universal).

In how far data managed in PANGEA can easily be integrated with the OCP remains to be assessed, based on a technical review.

2.3 GeoNetwork

GeoNetwork is a catalogue application to manage spatially referenced resources, vector and grid. It provides powerful metadata editing and search functions as well as an interactive web map viewer. It is currently used in numerous Spatial Data Infrastructure (SDI) initiatives across the world, for example ISRIC's soil datahub[20] or FAO's Geonetwork[21]. Entry of new metadata, however, will require a login (which may be restricted to specific communities).

GeoNetwork provides an easy to use web interface to search geospatial data across multiple catalogues. The search functionality provides full-text search as well as faceted search on keywords, resource types, organizations, scale, etc. Users can easily refine the search and quickly get to the records of interest.

GeoSpatial layers, but also services, maps or even non-geographic datasets can be described in the catalogue. Users can easily navigate across records and find sources or services publishing a dataset. Metadata on reports or papers that describe the methodology/study can be presented, with online access to scanned sources (e.g. PDFs or via DOI).

The interactive map viewer of GeoNetwork is based on OpenLayers 3. It is able to access to OGC[22] services (WMS, WMTS) and standards (KML, OWS). Connected to the catalogue, users can easily find new services, layers and even

[19] https://wiki.pangaea.de/wiki/Thesaurus
[20] http://data.isric.org/geonetwork/
[21] http://www.fao.org/geonetwork/srv/en/main.home
[22] http://www.opengeospatial.org

dynamic maps to combine them together; this functionality is not provided by Dataverse or PANGEA. User maps can be annotated, printed, and shared with others.

The Geonetwork editor also provides multilingual metadata editing, a validation system (e.g. against INSPIRE[23] requirements), suggestions to improve metadata quality, geo-publication of layers to publish geodata layers in OGC services (e.g. GeoServer).

GeoNetwork allows harvesting[24] from many sources including: OGC-CSW 2.0.2 ISO Profile, OAI-PMH, Z39.50 protocols, Thredds, Webdav, Web Accessible Folders, ESRI GeoPortal, Other GeoNetwork nodes. These should permit flexible communication with CIRCASA's OCP.

According to Lauregui[25], a RDF thesaurus can be uploaded in GeoNetwork; subsequently the corresponding keywords can be used to tag the metadata. These tags can then be used as search criteria. However, GeoNetwork does not use any of the more interesting semantic information from the thesaurus (e.g., it does not consider information about subclasses, synonyms, and other thesaurus relations). New Data Catalogue Vocabulary services under development at e.g. oSGEO[26] should be assessed on their merit for handling ontologies by a specialist or CIRCASA's TCP.

In how far data managed in PANGEA can be integrated with the OCP remains to be assessed, based on a technical review like for the other platforms under review.

2.4 GeoNode

GeoNode[27] is a web-based geospatial content management system. By combining information found in social networks with specialized geospatial tools, GeoNode makes it easy to explore, process, style, and share maps and geospatial data.

Spatial datasets can be imported and shared, all through a non-technical user interface. Features include powerful spatial search engine, federated OGC services, and metadata catalogue.

Geonodes are designed to be extended and modified, and can be integrated/harvested into existing platforms. GeoNode makes it easy to upload and manage geospatial data on the web. Any user can upload and make content available via standard OGC protocols such as Web Map Service (WMS) and Web Feature Service (WFS). Data is available for browsing, searching, styling, and processing to generate maps that can be shared publicly or restricted to specific users only. Figure 4 gives an example of a customised GeoNode instance as developed for the ICRAF Landscape Portal.

[23] https://inspire.ec.europa.eu/requirements-inspire-directive/64
[24] https://geonetwork-opensource.org/manuals/2.10.4/eng/users/managing_metadata/harvesting/index.html
[25] https://groups.google.com/forum/#!topic/geonode-users/AhjhEArAbZ8
[26] https://trac.osgeo.org/geonetwork/wiki/proposals/DCATandRDFServices
[27] http://geonode.org/

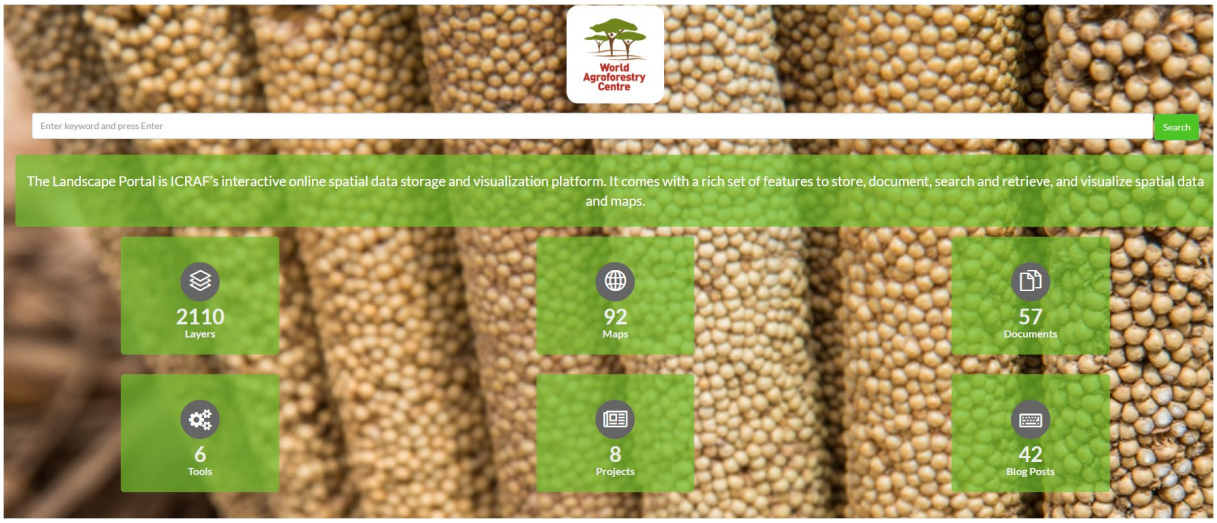


Figure 4. Example of customised GeoNode instance, ICRAF Landscape Portal.

Supported upload formats include shapefile, GeoTIFF, KML and CSV. In addition, it is possible to connect to existing external spatial databases and services.

Features of a Geonode include: publish raster, vector, and tabular data; manage metadata and associated documents; securely or publicly share data; versioned geospatial data editor.

GeoNode comes with helpful cartography tools for styling and composing maps graphically. These tools make it easy for anyone to assemble a web-based mapping application with functionality traditionally found in desktop GIS applications.

Users can gain enhanced interactivity with GIS-specific tools such as querying and measuring. GeoNode's documentation is provided at http://docs.geonode.org/en/master/.

GeoNode provides a JSON API which currently supports the GET method. The API is also used as main search engine.

Like for the other platforms under review, a technical review is needed to assess in how far data managed in GeoNode can be integrated with the OCP.

4 Ontologies

Although not referring to any specific platform for handling metadata, some observations are made here concerning the need for a consistent ontology as without it user management requirements cannot be met (CIRCASA/ISRIC 2018). Ontology refers to a 'set of concepts and categories in a subject area or domain that shows their properties and the relations between them'.

In the case of CIRCASA, data in various formats will be sourced from different organizations using different information technology facilities, posing a serious challenge as to how to integrate the data into a workable system so that diverse user groups can query the system unambiguously. One of the most important requirements of such an integration is that the data semantics are consistent among the different phases of the process. For this, a consistent ontology should be utilized by all the information systems of the different phases.

The need for a 'solid, unambiguous' ontology has been illustrated in a test case for a GeoNode for Malawi[28]. In this example it appeared that the ontology of the site did not match users' mental models for the data's ontology (subject categories that show their properties and inter-category relationships).

Every discipline creates ontologies to limit complexity and organize information into data and knowledge. As new ontologies are made, their use hopefully improves problem solving within that domain (Bonacin *et al.* 2016; Hu *et al.* 2011; Su *et al.* 2012; Zheng *et al.* 2012). Accessing research papers within every field is made easier when experts from different countries/institutions maintain a controlled vocabulary of jargon between each of their languages. In view of the above, choice of an agreed ontology for CIRCASA is considered critical irrespective of the platforms selected, this in accord with decisions of the TCP. Preferably, the adopted ontology should be compatible (i.e. interoperable) with procedures adopted for say the 4p1000 OCP or the CGIAR big data platform[29].

INRA adopted the semantic web's practices and standards (RDF, SKOS, OWL, SPARQL) to enable the methodological and technical practices needed by INRA's scientists to standardize, document and publish the vocabularies created in their projects (Jonquet *et al.* 2018).

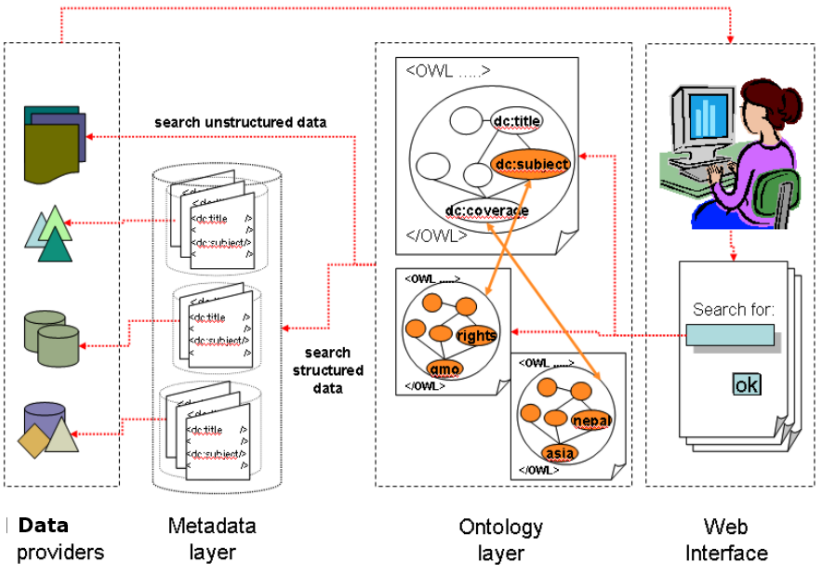


Figure 5. A consistent metadata standard and ontology is needed to exchange data from multiple sources[28].

[28] http://geonode.org/docs/GeoNode_Homepage_UabilityAnalysisReport_March2017.pdf
[29] https://bigdata.cgiar.org/ontologies/

Collaborative efforts include AgroPortal, a vocabulary and ontology repository for agronomy, food, plant sciences and biodiversity (Jonquet *et al.* 2018). This platform already hosts 64 ontologies from varied areas related to agriculture, including AGROVOC[30] and the AnaEE thesaurus of INRA. According to Jonquet et al. (2018), the AgroPortal specifically satisfies requirements of the agronomy community in terms of ontology formats (e.g., SKOS vocabularies and trait dictionaries) and supported features (offering detailed metadata and advanced annotation capabilities). Through AgroPortal[31], ontology-based projects can be browsed/searched, new projects created, and different ontologies mapped. As illustrated in Figure 5, a consistent metadata standard and ontology are needed to effectively exchange data from multiple sources[28].

5 Concluding remarks and next steps

Based on this short, non-technical, evaluation it appears that none of the platforms under review fulfil the comprehensive specification of requirements for the KIS as presented in DocD1.2 (CIRCASA/ISRIC 2018). In particular, some of the 'must have' requirements on the MoSCoW scale, such as OGC compliance, are not met by some platforms.

The KIS for CIRCASA will have to be able to handle diverse data types (publications and documents, point data, polygon data, grid data), use an unambiguous search system, have defined user and user profiles/roles, provide dataset access within user profiles, and follow consistent standards to permit seamless integration with the upcoming OCP. Developing a comprehensive, 'state-of-the-art' system meeting at least all 'must have' and 'should have' requirements would require quite some software-engineering which, realistically, may be beyond ISRIC's niche.

A customised Dataverse implementation may seem a pragmatic solution for CIRCASA's knowledge information system, complemented with GeoNetwork as a versatile, metadata catalogue for handling a broader range of spatial data. Being a collaborative H2020 project, the platform(s) should be customised with the CIRCASA logo (i.e. not any particular organisation), with the understanding though that the Dataverse will need to be sustained by a host institution after completion of the project.

A perceived advantage of these two platforms is that they are open-source, hence maintained by a wide and dynamic community. However, only GeoNetwork follows OGC standards as recommended in DocD1.2.

For ease of use, preference may be for platforms that use English as the recommended language or have multi-lingual capability.

CIRCASA partners should make use of, or build upon, an existing, multi-lingual ontology such as maintained by AgroPortal, in accord with TCP's decision on the matter.

Use of Creative Commons licences is advocated for the various data sources; this is in line with the recommendations of the ICSU World Data System.

[30] http://aims.fao.org/vest-registry/vocabularies/agrovoc
[31] http://agroportal.lirmm.fr/

A strategic choice has to be made by the CIRCASA management team on how to proceed in relation to deliverable D1.5 (operational pilot knowledge information system), with a possible re-allocation of tasks, and the integration into the OCP. A practical solution would be to adopt Dataverse for the KIS and contract a software-engineer to develop an OGC-compliant integration or plug-in to handle a wider range of spatial data.

Acknowledgements

This document was prepared in the framework of the EU 2020 CIRCASA project. I thank Luis de Sousa for constructive comments, and discussion, on an earlier version of the document.

The findings were discussed during a Skype meeting with representatives of INRA and ISRIC (25 October 2018) to identify the way forward.

References

Bonacin R, Nabuco OF and Pierozzi Junior I 2016. Ontology models of the impacts of agriculture and climate changes on water resources: Scenarios on interoperability and information recovery. *Future Generation Computer Systems* 54, 423-434. https://doi.org/10.1016/j.future.2015.04.010

CIRCASA/ISRIC 2018. *Knowledge Information System Requirements Specification (deliverable D1.2; Compiled by Luís de Sousa/ISRIC)*, Coordination of International Research Cooperation on Soil Carbon Sequestration in Agriculture (CIRCASA), Wageningen. https://www.circasa-project.eu/content/download/3660/35410/version/1/file/D1.2.pdf

FAO and ITPS 2015. *Status of the world's soil resources (SWSR) - Main report*, Food and Agriculture Organization of the United Nations and Intergovernmental Technical Panel on Soils, Rome, 650 p. http://www.fao.org/3/a-i5199e.pdf

Harden JW, Hugelius G, Ahlström A, Blankinship JC, Bond-Lamberty B, Lawrence CR, Loisel J, Malhotra A, Jackson RB, Ogle S, Phillips C, Ryals R, Todd-Brown K, Vargas R, Vergara SE, Cotrufo MF, Keiluweit M, Heckman KA, Crow SE, Silver WL, DeLonge M and Nave LE 2017. Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter. *Global Change Biology* 24, e705-e718. https://doi.org/10.1111/gcb.13896

Hu S, Wang H, She C and Wang J 2011. AgOnt: Ontology for Agriculture Internet of Things. Computer and Computing Technologies in Agriculture IV. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 131-137.

Jonquet C, Toulet A, Arnaud E, Aubin S, Dzalé Yeumo E, Emonet V, Graybeal J, Laporte M-A, Musen MA, Pesce V and Larmande P 2018. AgroPortal: A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture* 144, 126-143. https://doi.org/10.1016/j.compag.2017.10.012

Madalli DP 2015. Thematic harvesting of agricultural resources from generic repositories. *Information Processing in Agriculture* 2, 93-100. https://doi.org/10.1016/j.inpa.2015.05.002

Soussana J-F, Lutfalla S, Ehrhardt F, Rosenstock T, Lamanna C, Havlík P, Richards M, Wollenberg E, Chotte J-L, Torquebiau E, Ciais P, Smith P and Lal R 2017. Matching policy and science: Rationale for the '4 per 1000 - soils for food security and climate' initiative. *Soil and Tillage Research*. https://doi.org/10.1016/j.still.2017.12.002

Su X-l, Li J, Cui Y-p, Meng X-x and Wang Y-q 2012. Review on the Work of Agriculture Ontology Research Group. *Journal of Integrative Agriculture* 11, 720-730. https://doi.org/10.1016/S2095-3119(12)60061-6

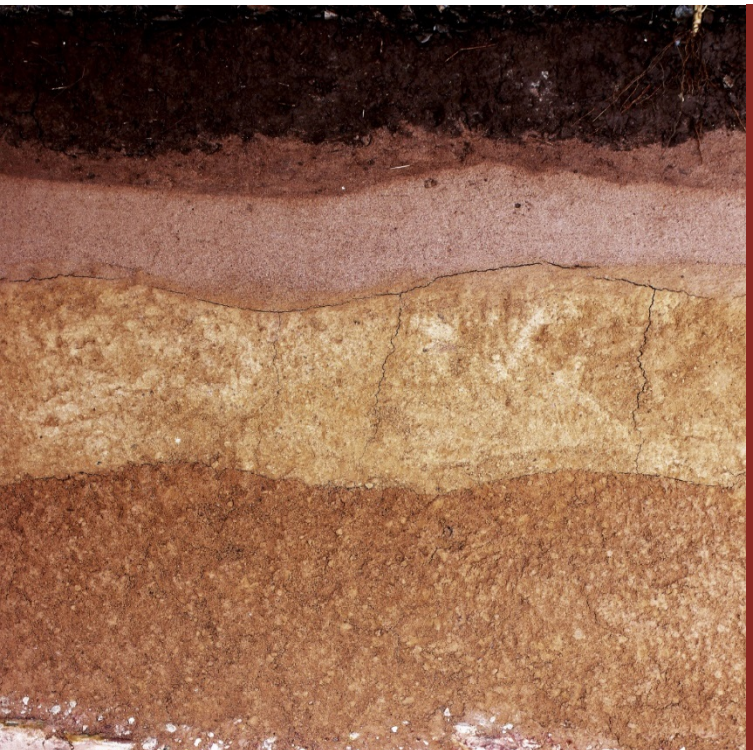
Sulman BN, Moore JAM, Abramoff R, Averill C, Kivlin S, Georgiou K, Sridhar B, Hartman MD, Wang G, Wieder WR, Bradford MA, Luo Y, Mayes MA, Morrison E, Riley WJ, Salazar A, Schimel JP, Tang J and Classen AT 2018. Multiple models and experiments underscore large uncertainty in soil carbon dynamics. *Biogeochemistry*. https://doi.org/10.1007/s10533-018-0509-z doi:10.1007/s10533-018-0509-z

UNCCD 2017. *The Global Land Outlook (First Edition)*, United Nations Convention to Combat Desertification, Bonn, 336 p. https://knowledge.unccd.int/sites/default/files/2018-06/GLO%20English_Full_Report_rev1.pdf

UNEP 2012. The benefits of soil carbon - managing soils for multiple, economic, societal and environmental benefits, *UNEP Yearbook - Emerging issues in our global environment 2012*. United Nations Environmental Programme, Nairobi, pp 19-33

Zheng Y-l, He Q-y, Qian P and Li Z 2012. Construction of the Ontology-Based Agricultural Knowledge Management System. *Journal of Integrative Agriculture* 11, 700-709. https://doi.org/10.1016/S2095-3119(12)60059-8

Together with our partners we produce, gather, compile and serve quality-assured soil information at global, national and regional levels. We stimulate the use of this information to address global challenges through capacity building, awareness raising and direct cooperation with users and clients.



www.isric.org