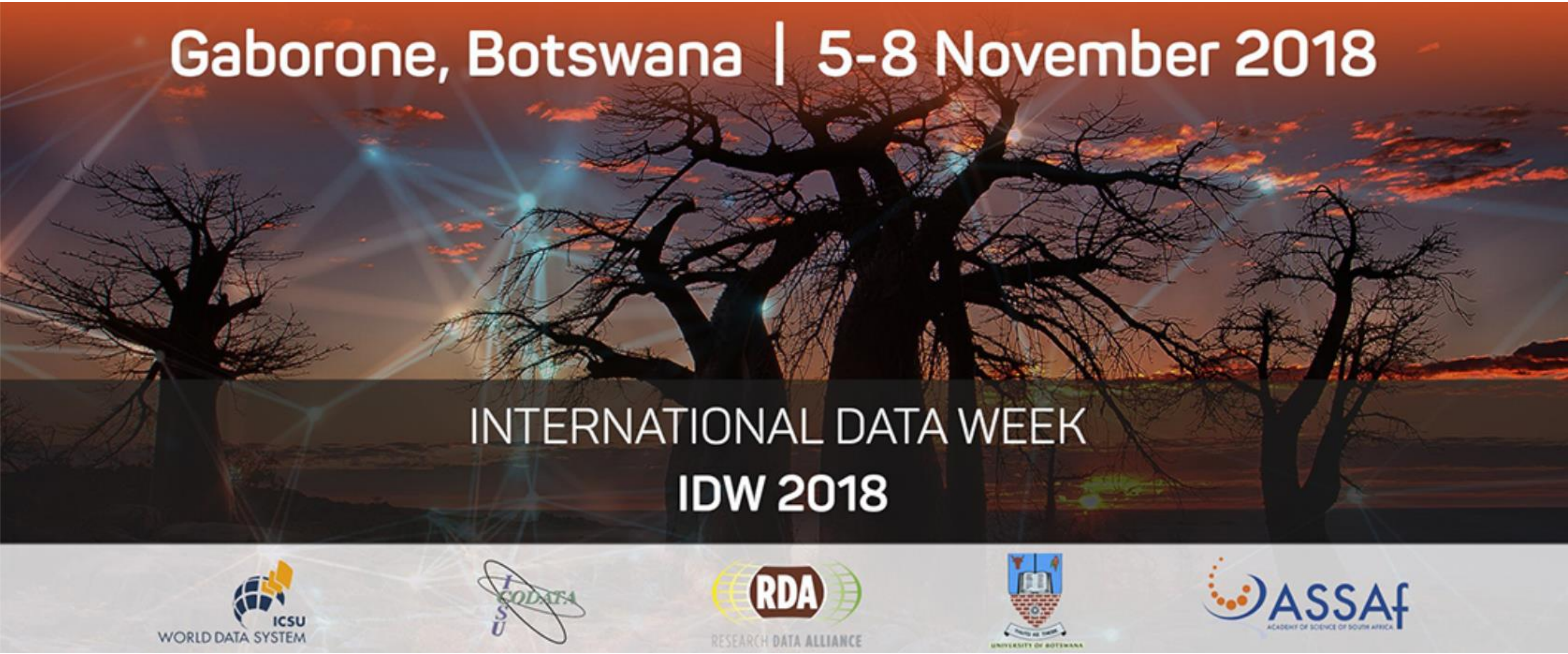
International Center for Tropical Agriculture
Since 1967 Science to cultivate change

Research Data Publishing

15 Nov 2018, CIAT Seminar Series, CIAT HQ

Leroy Mwanzia
Chief Data Officer
l.mwanzia@cgiar.org

CGIAR
A CGIAR Research Center



Gaborone, Botswana | 5-8 November 2018

INTERNATIONAL DATA WEEK

IDW 2018

RDA

RESEARCH DATA ALLIANCE

RESEARCH DATA SHARING WITHOUT BARRIERS

www.rd-alliance.org

RDA Interest Groups and Working Groups

RDA/WDS Publishing Data IG

RDA Interest Groups and Working Groups

RDA/WDS Publishing Data IG

Domain Repositories Interest Group

Preservation e-Infrastructure IG

Preservation Tools, Techniques, and Policies

Sharing Rewards and Credit (SHARC) IG

Data Discovery Paradigms IG

Repository Platforms for Research Data IG

Reproducibility IG

Research Data Provenance IG

Sharing Rewards and Credit (SHARC) IG

RDA/NISO Privacy Implications of Research Data Sets IG

FAIRSharing Registry: connecting data policies, standards & databases WG

Metadata IG

Data Usage Metrics WG

Data Versioning WG

FAIR Data Maturity Model WG

Data Citation WG

RDA/WDS Publishing Data Workflows WG

Research Data Collections WG

Data Description Registry Interoperability (DDRI) WG

Research Data Repository Interoperability WG

Research Data Publishing

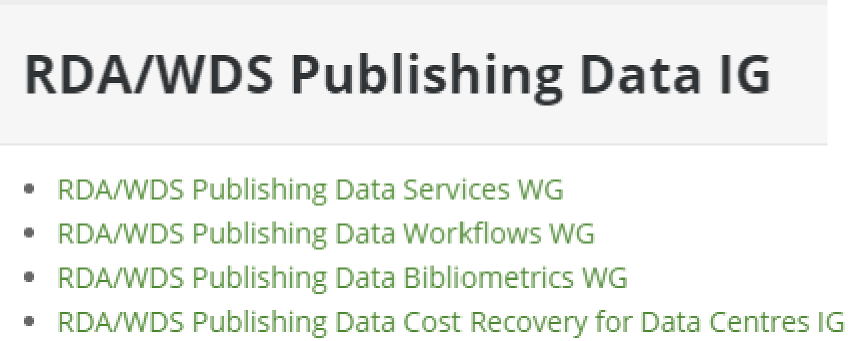
"The release of research **data**, associated **metadata**, accompanying **documentation**, and **software code** (in cases where the raw data have been processed or manipulated) for re-use and analysis in such a manner that they can be discovered on the Web and referred to in a *unique* and *persistent* way."

(Austin et al 2015)

Austin, C. C., Bloom, T., Dallmeier-Tiessen, S., Khodiyar, V., Murphy, F., Nurnberger, A., … Whyte, A. (2015). Key components of data publishing: Using current best practices to develop a reference model for data publishing. http://doi.org/10.5281/zenodo.34542

Data as First Class Research Product

- Data should be considered legitimate, citable products of research.[1]
- That can be validated, preserved, cited and credit[2].



RDA/WDS Publishing Data IG

- RDA/WDS Publishing Data Services WG
- RDA/WDS Publishing Data Workflows WG
- RDA/WDS Publishing Data Bibliometrics WG
- RDA/WDS Publishing Data Cost Recovery for Data Centres IG

RDA RESEARCH DATA ALLIANCE

WEB OF SCIENCE™
THOMSON REUTERS
ABOUT | PRODUCTS & TOOLS | BENEFITS & RESOURCES | TRAINING & SUPPORT | NEWS & EVENTS | CONTACT US
Site Search
SEARCH
Products and Tools › Multidisciplinary › Data Citation Index
What is it? VIEW VIDEO
THE DATA CITATION INDEX℠ ON WEB OF SCIENCE

THE DATA CITATION INDEX℠

CONNECTING THE DATA TO THE RESEARCH IT INFORMS

FORCE11 The Future of Research Communications and e-Scholarship

DC1 Data Citation Principles

Data Citation Principles

- Importance
- Credit and Attribution
- Evidence
- Unique Identification
- Access
- Persistence
- Specificity and Verifiability
- Interoperability and Flexibility

1. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014 [https://www.force11.org/group/joint-declaration-data-citation-principles-final
2. Kratz J and Strasser C. Data publication consensus and controversies [version 1; referees: 1 approved with reservations]. *F1000Research* 2014, **3**:94 (doi: 10.12688/f1000research.3979.1)

Why Publish Research Data?

Sharing data wasn't cool, but neither were we – how WorldClim changed my life

by Andy Jarvis | Oct 26, 2017

https://blog.ciat.cgiar.org/sharing-data-wasnt-cool-but-neither-were-we-how-worldclim-changed-my-life/

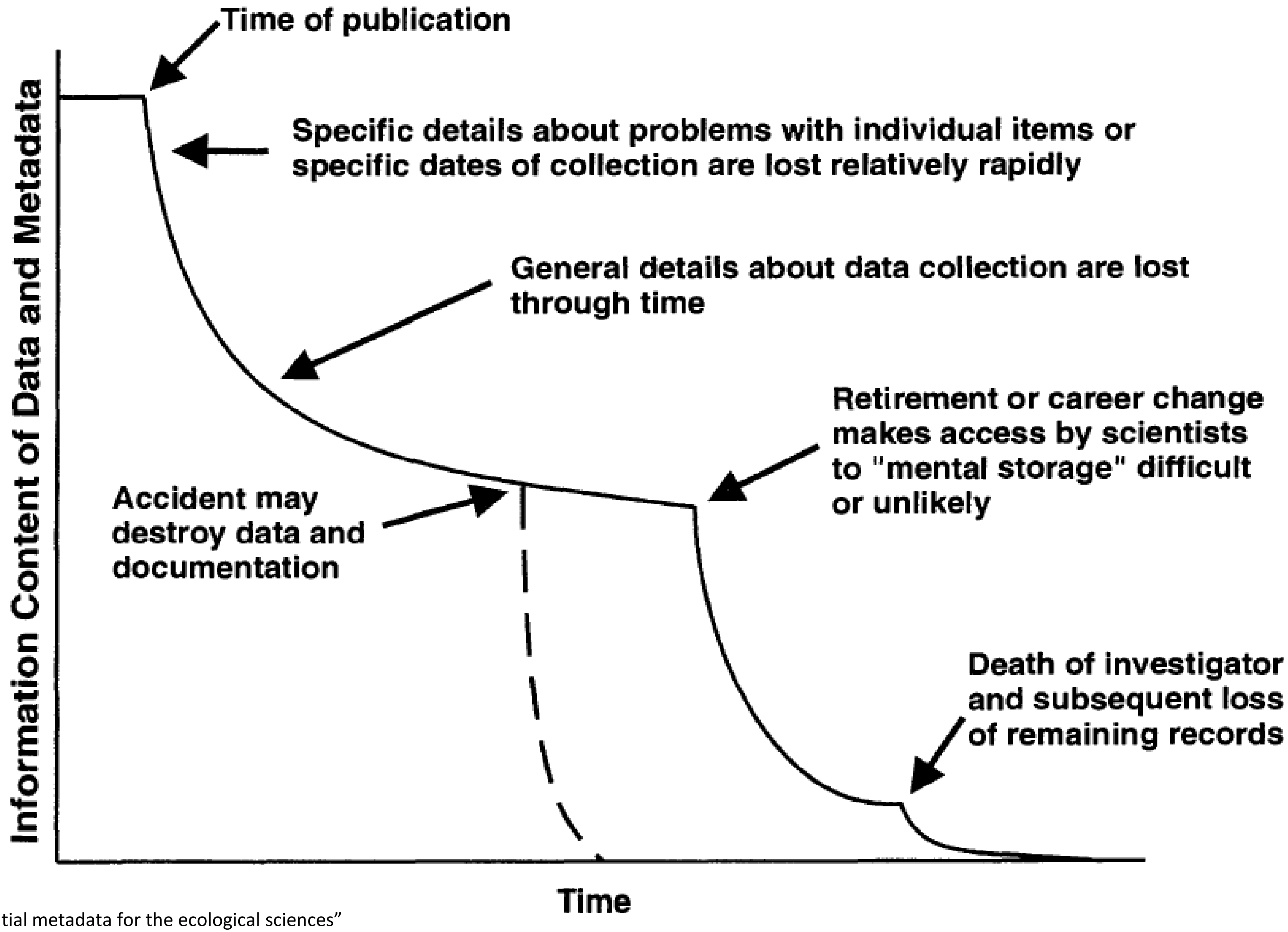
Why Publish Research Data

Recognition & attribution: Can provide a direct credit to the researcher and institution.

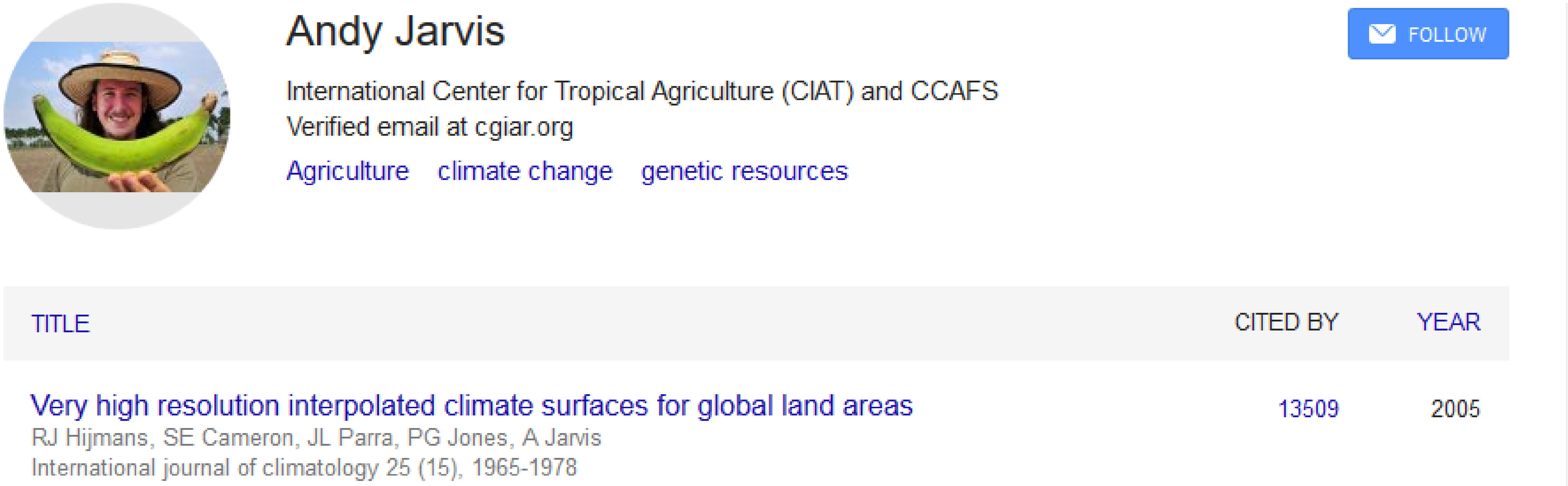
Why Publish Research Data

Increases the impact and visibility of research

Information entropy



Michener et al. 1997 “Nongeospatial metadata for the ecological sciences”



Andy Jarvis

International Center for Tropical Agriculture (CIAT) and CCAFS
Verified email at cgiar.org

Agriculture climate change genetic resources

TITLE	CITED BY	YEAR
Very high resolution interpolated climate surfaces for global land areas RJ Hijmans, SE Cameron, JL Parra, PG Jones, A Jarvis International journal of climatology 25 (15), 1965-1978	13509	2005

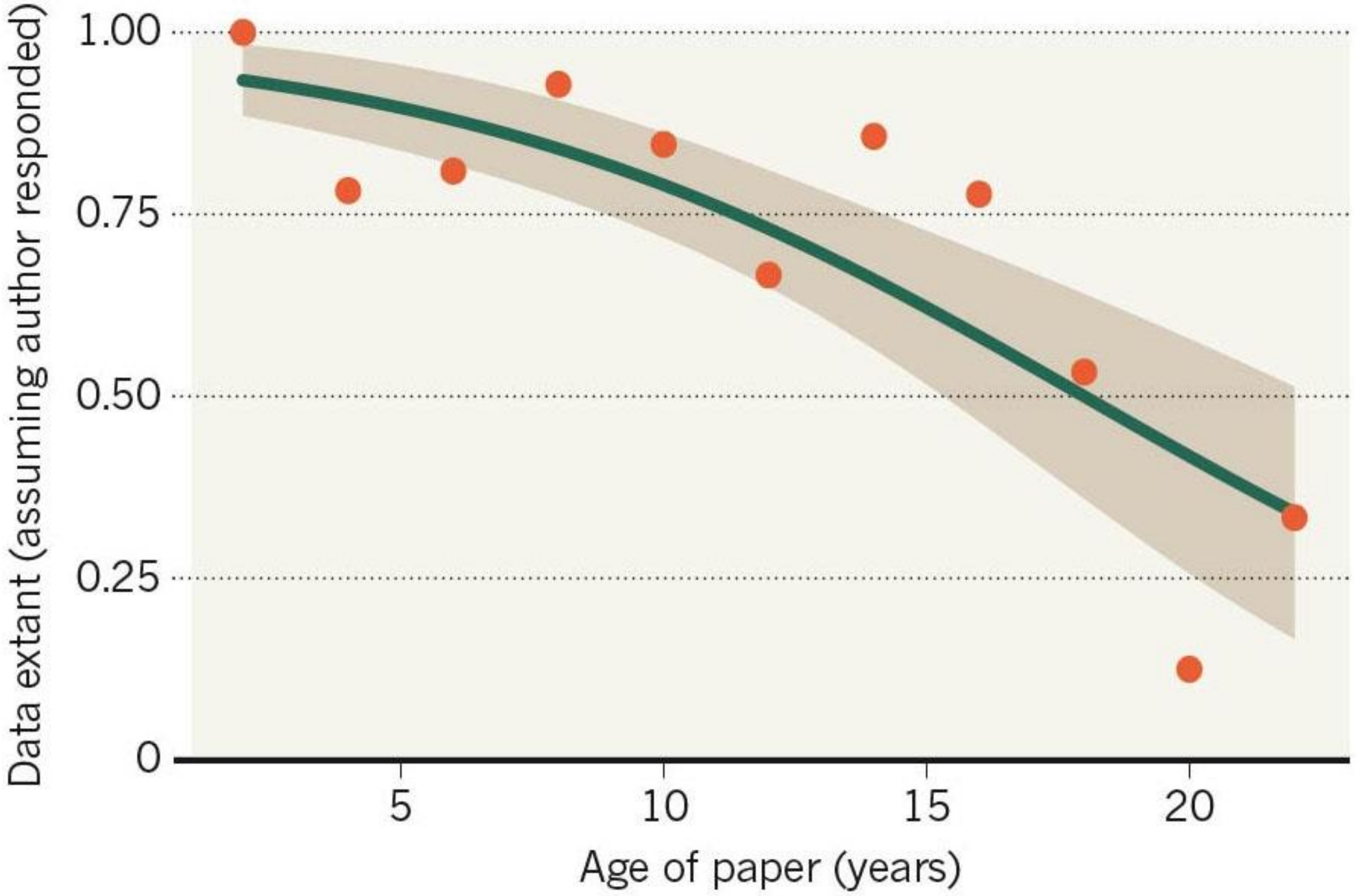
Why Publish Research Data

Facilitating science: discovery & access reinforces open scientific enquiry: Reproducibility and transparency.

Icons from www.flaticon.com licensed CC 3.0 BY

MISSING DATA

As research articles age, the odds of their raw data being extant drop dramatically.

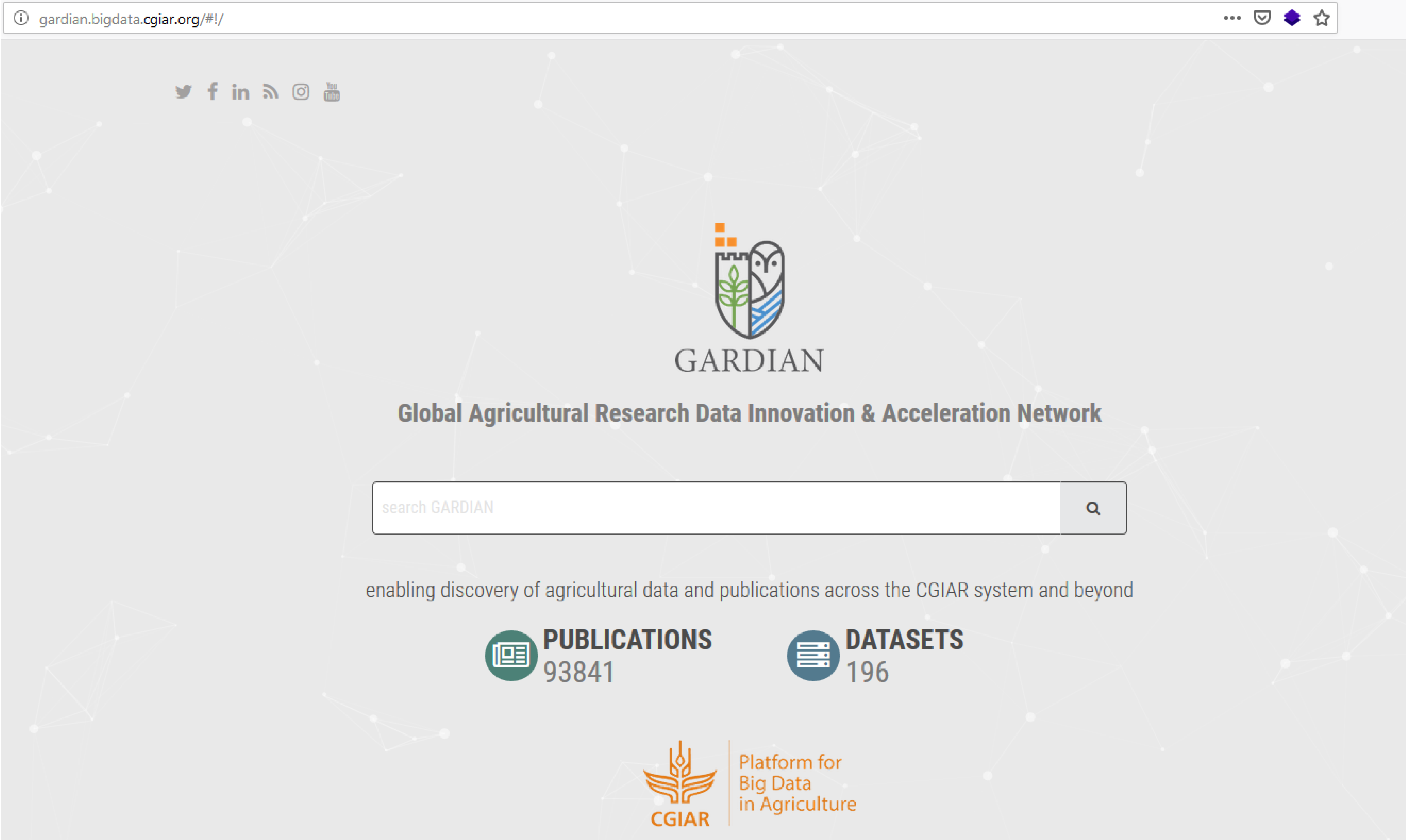


(2014) Vines et al. "Availability of research data was strongly affected by article's age when no policy is in place"

Reworked image credit: Kristin Briney

Why Publish Research Data

Promotes the research & demonstrates use and relevance of the research



gardian.bigdata.cgiar.org/#!/

GARDIAN

Global Agricultural Research Data Innovation & Acceleration Network

search GARDIAN

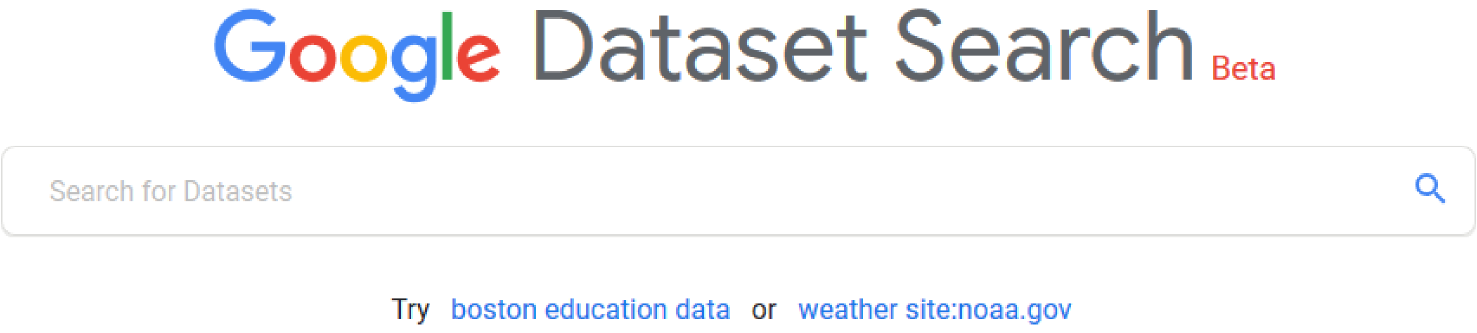
enabling discovery of agricultural data and publications across the CGIAR system and beyond

PUBLICATIONS
93841

DATASETS
196

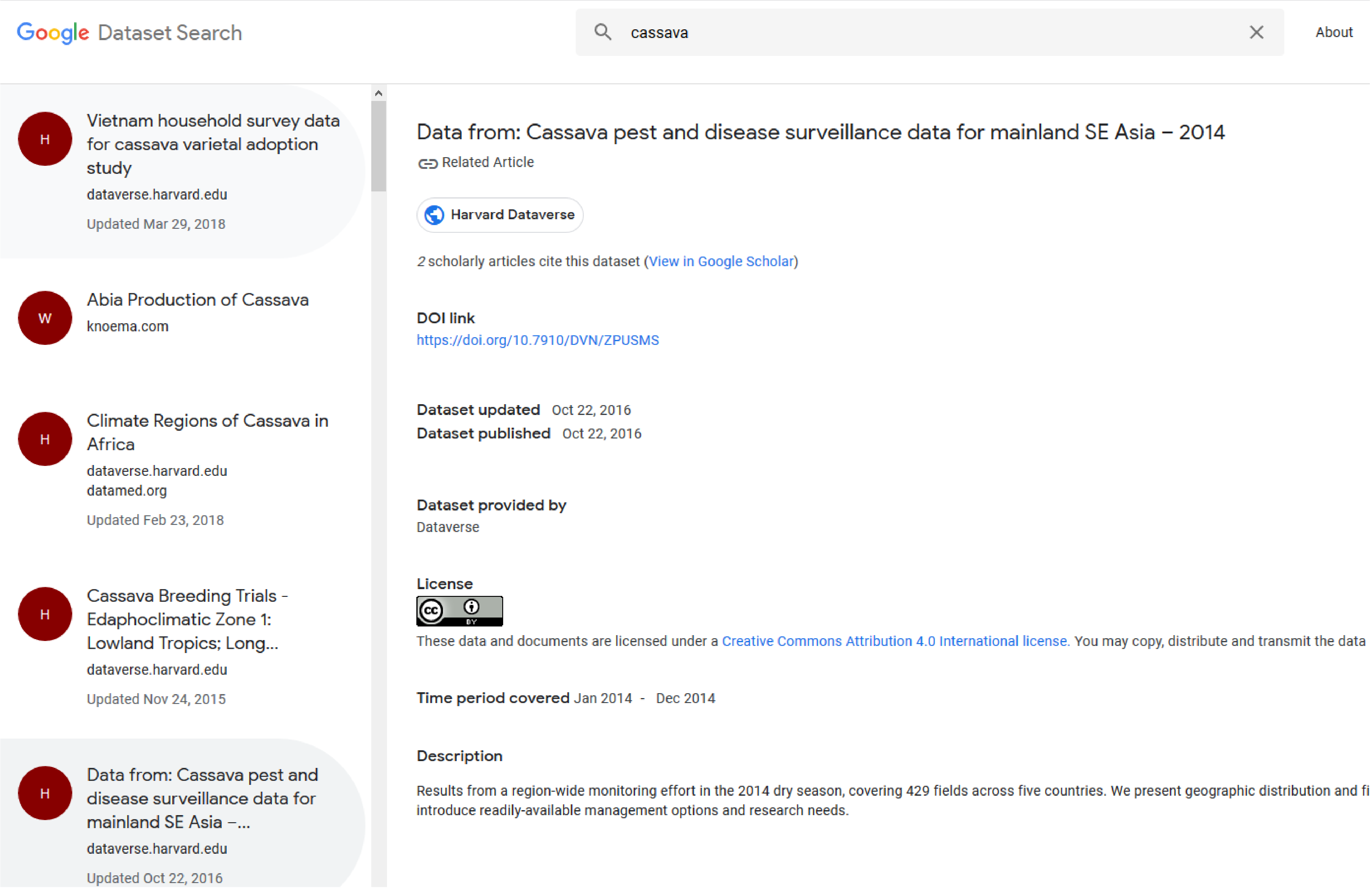
CGIAR

Platform for
Big Data
in Agriculture



Google Dataset Search Beta

Try boston education data or weather site:noaa.gov



Vietnam household survey data for cassava varietal adoption study
dataverse.harvard.edu
Updated Mar 29, 2018

Abia Production of Cassava
knoema.com

Climate Regions of Cassava in Africa
dataverse.harvard.edu
datamed.org
Updated Feb 23, 2018

Cassava Breeding Trials - Edaphoclimatic Zone 1: Lowland Tropics; Long...
dataverse.harvard.edu
Updated Nov 24, 2015

Data from: Cassava pest and disease surveillance data for mainland SE Asia –...
dataverse.harvard.edu
Updated Oct 22, 2016

Data from: Cassava pest and disease surveillance data for mainland SE Asia – 2014

Related Article

Harvard Dataverse

2 scholarly articles cite this dataset (View in Google Scholar)

DOI link
https://doi.org/10.7910/DVN/ZPUSMS

Dataset updated Oct 22, 2016
Dataset published Oct 22, 2016

Dataset provided by
Dataverse

License

These data and documents are licensed under a Creative Commons Attribution 4.0 International license. You may copy, distribute and transmit the data

Time period covered Jan 2014 - Dec 2014

Description

Results from a region-wide monitoring effort in the 2014 dry season, covering 429 fields across five countries. We present geographic distribution and fi
introduce readily-available management options and research needs.

Why Publish Research Data

Reduces the cost of duplication – Increase efficiency

Barriers to Publishing Data?

Leroy, I am too busy with this CRP reports. I don’t have time!

He published with my data and did not acknowledge me in any way!

I shared data with them and they published on the exact same topic before I did. Shame on them!

I would share my data but I don’t think my data is clean enough!

Barriers to Publishing Data

- Lack of attribution
- Data citation practices not well known and not universally agreed
- Lack of incentives and rewards
- Data quality issues
- Lack of data sharing culture

Types of Data to Publish

- Primary data used in the production of a publication.

- Unpublished datasets that span an entire research project and that are described by:
  - Materials and methods
  - Proper documentation including a clear description of the variables, data acquisition tools, software code if the data was transformed from its raw format.

Before you publish: Prepare Data

Ensure dataset is cleaned, verified for correctness and fitness for use (Keep the raw dataset)

Ensure data is well structured

Ensure data is well documented

Ensure you have considered privacy, confidentiality and security related issues

Use reusable file formats

FAIR Data

The key consideration when selecting where to publish data is to ensure that data will at the end adhere to the FAIR data principles.

Findable

- unique identifier, rich metadata, indexed

Accessible

- Retrievable by identifier, by: standard, open, free, authenticatable protocols

Interoperable

- uses formal, shared & applicable knowledge representation, human readable/machine readable, FAIR vocabularies

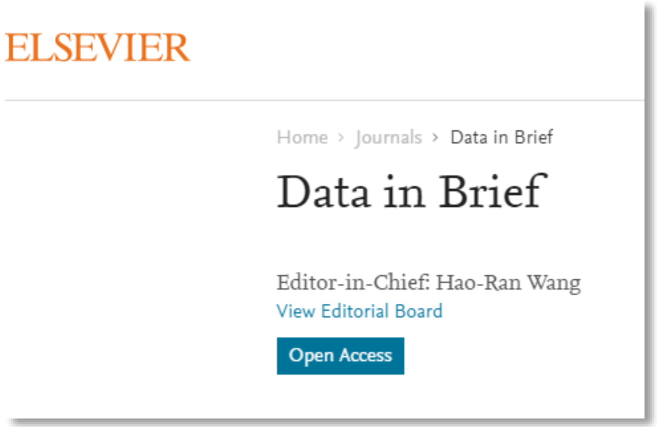
Reusable

- Provenance, data usage license, domain relevance standards

Where: Peer Reviewed Data Journals

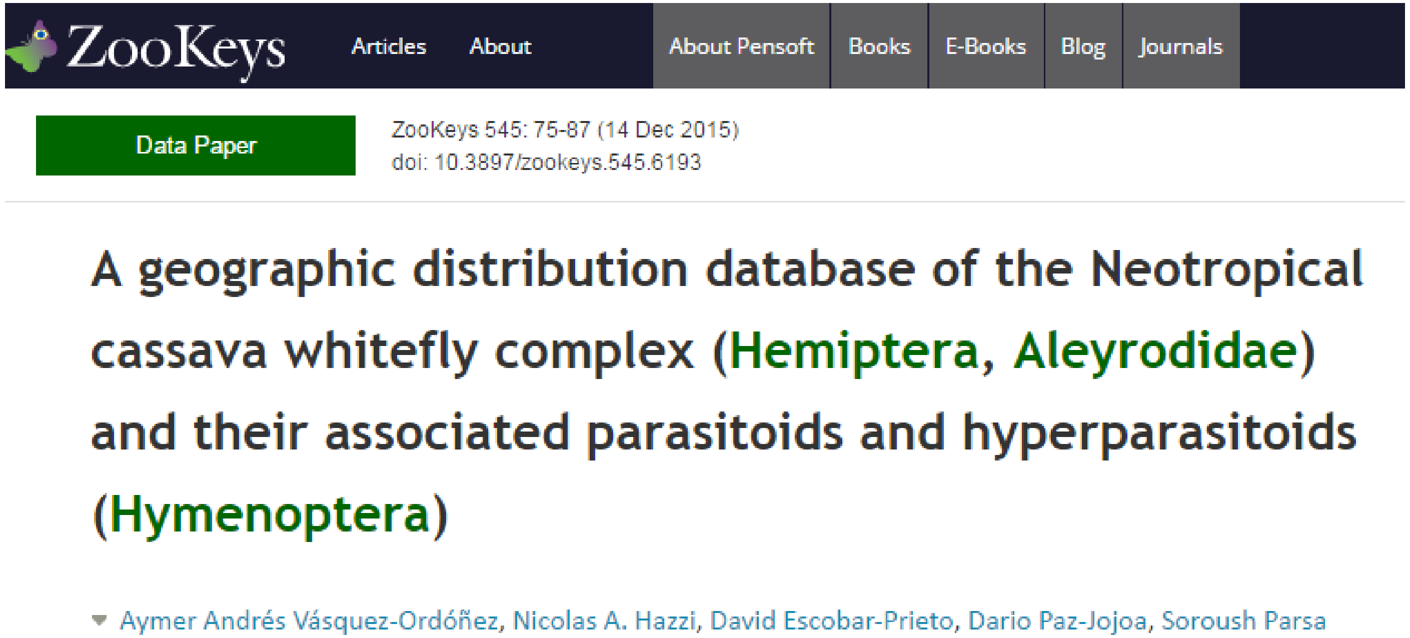
& Introductory Earth Sciences > Geoscience Data Journal

Geoscience Data Journal



Data in Brief

Editor-in-Chief: Hao-Ran Wang



ZooKeys 545: 75-87 (14 Dec 2015)
doi: 10.3897/zookeys.545.6193

A geographic distribution database of the Neotropical cassava whitefly complex (Hemiptera, Aleyrodidae) and their associated parasitoids and hyperparasitoids (Hymenoptera)

Aymer Andrés Vásquez-Ordóñez, Nicolas A. Hazzi, David Escobar-Prieto, Dario Paz-Jojoa, Soroush Parsa

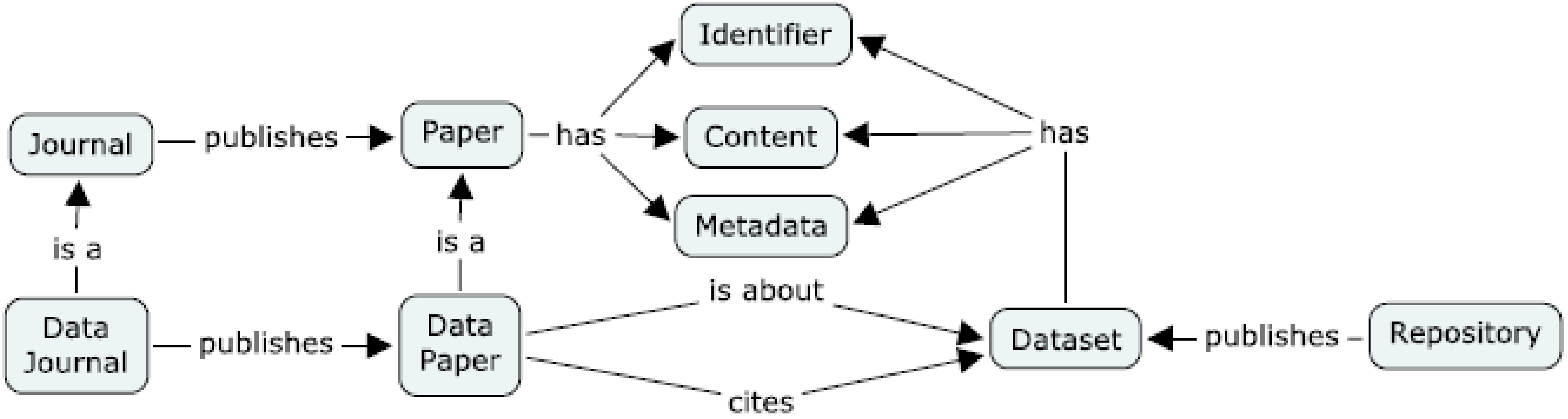
Peer Reviewed Data Paper

"A scholarly publication of searchable metadata document describing a particular on-line accessible dataset, or a group of datasets, published in accordance to standard academic practices" [1]

The primary purpose of a data paper is to describe data and the circumstances of their collection, rather than to report hypotheses and conclusions.

[1] Chavan, V. and Penev, L., 2011. The data paper: a mechanism to incentivize data publishing in biodiversity science. *BMC bioinformatics*, *12*(15), p.S2.

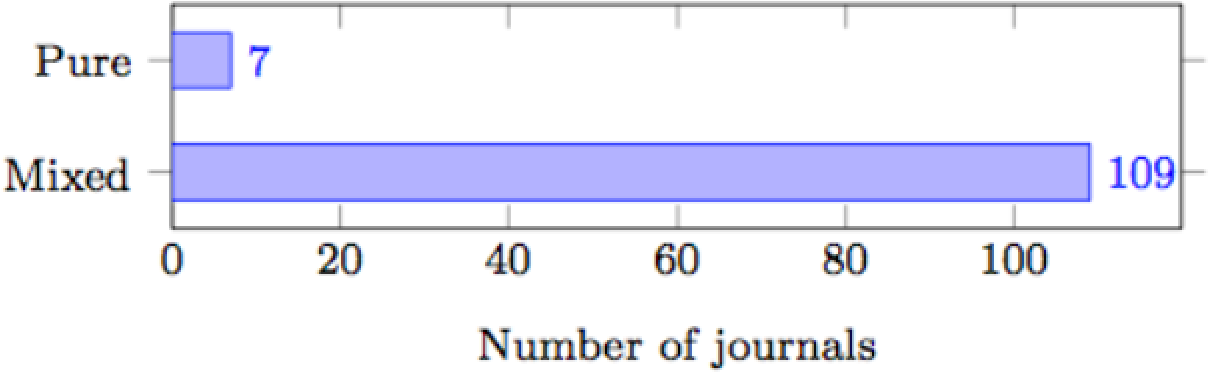
Peer Reviewed Data Paper Concept Map



Candela, L., Castelli, D., Manghi, P. and Tani, A. (2015), Data journals: A survey. J Assn Inf Sci Tec, 66: 1747–1762. doi:10.1002/asi.23358

Nature of Data Publishing Journals

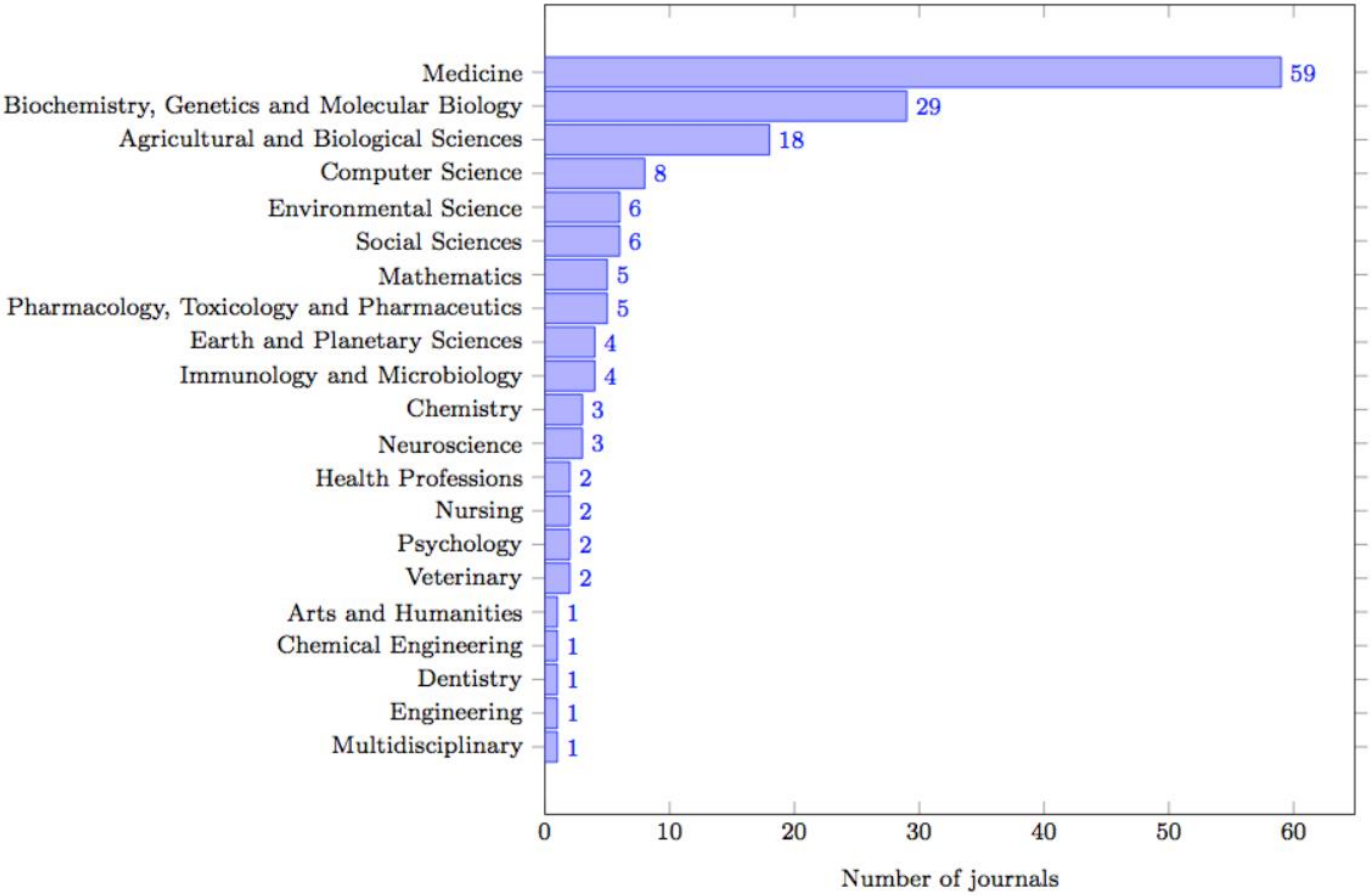
- Pure – Journal only publishes data papers e.g. Scientific Data from Nature and Data in Brief from Elsevier
- Mixed – Journal publishes different types of scientific papers including data papers. e.g. PLOS One and F1000 Research.



2015

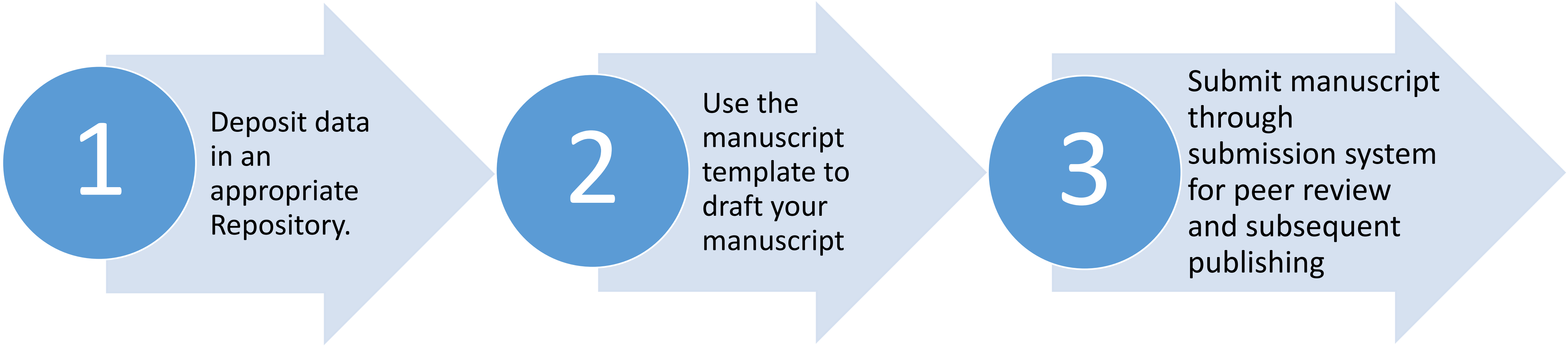
Candela, L., Castelli, D., Manghi, P. and Tani, A. (2015), Data journals: A survey. J Assn Inf Sci Tec, 66: 1747–1762. doi:10.1002/asi.23358

Number of Journals by Subject (2015)



Candela, L., Castelli, D., Manghi, P. and Tani, A. (2015), Data journals: A survey. J Assn Inf Sci Tec, 66: 1747–1762. doi:10.1002/asi.23358

General Peer Reviewed Publishing Process



1 Deposit data in an appropriate Repository.

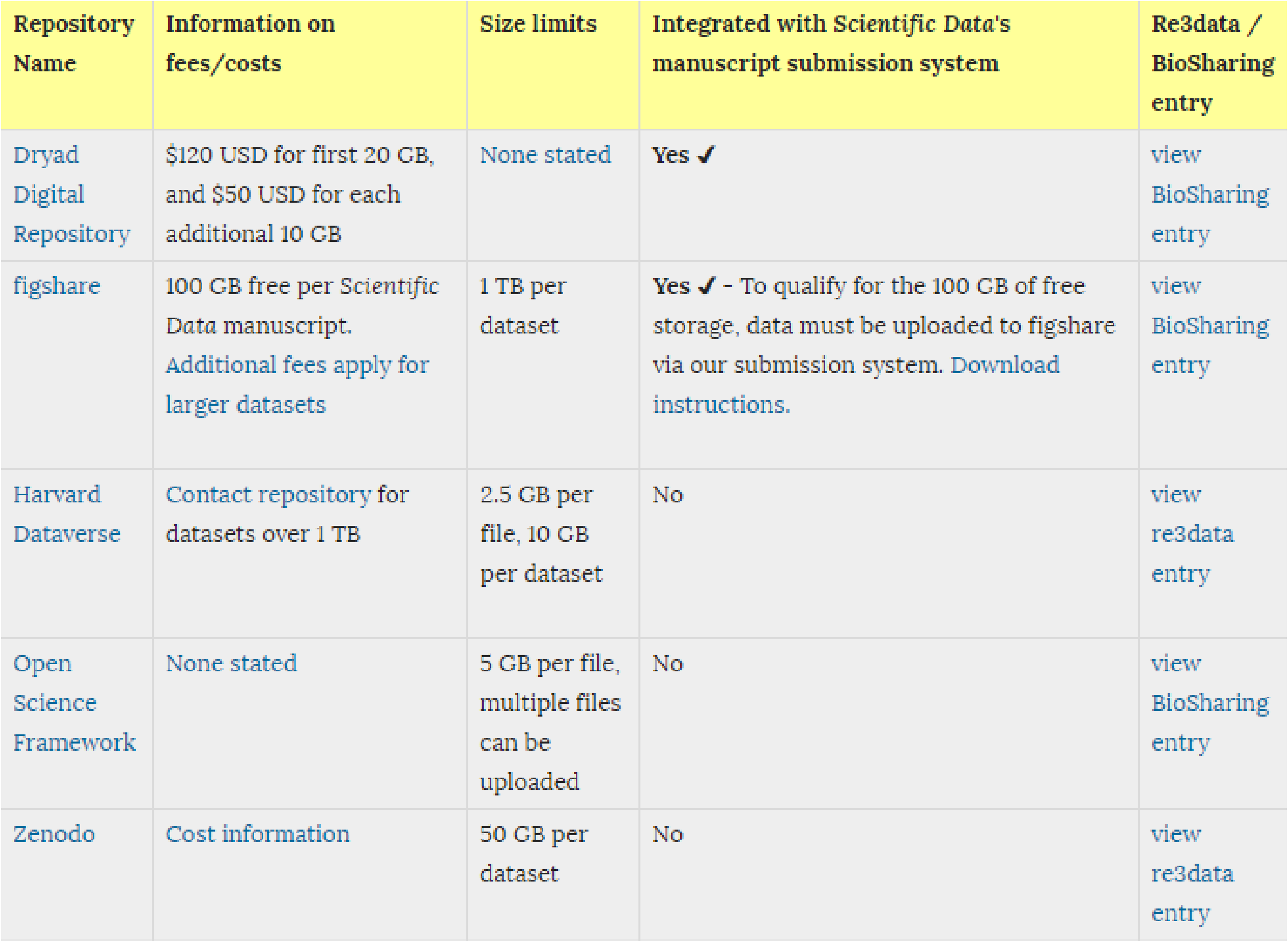
2 Use the manuscript template to draft your manuscript

3 Submit manuscript through submission system for peer review and subsequent publishing

Contents of the Data Paper

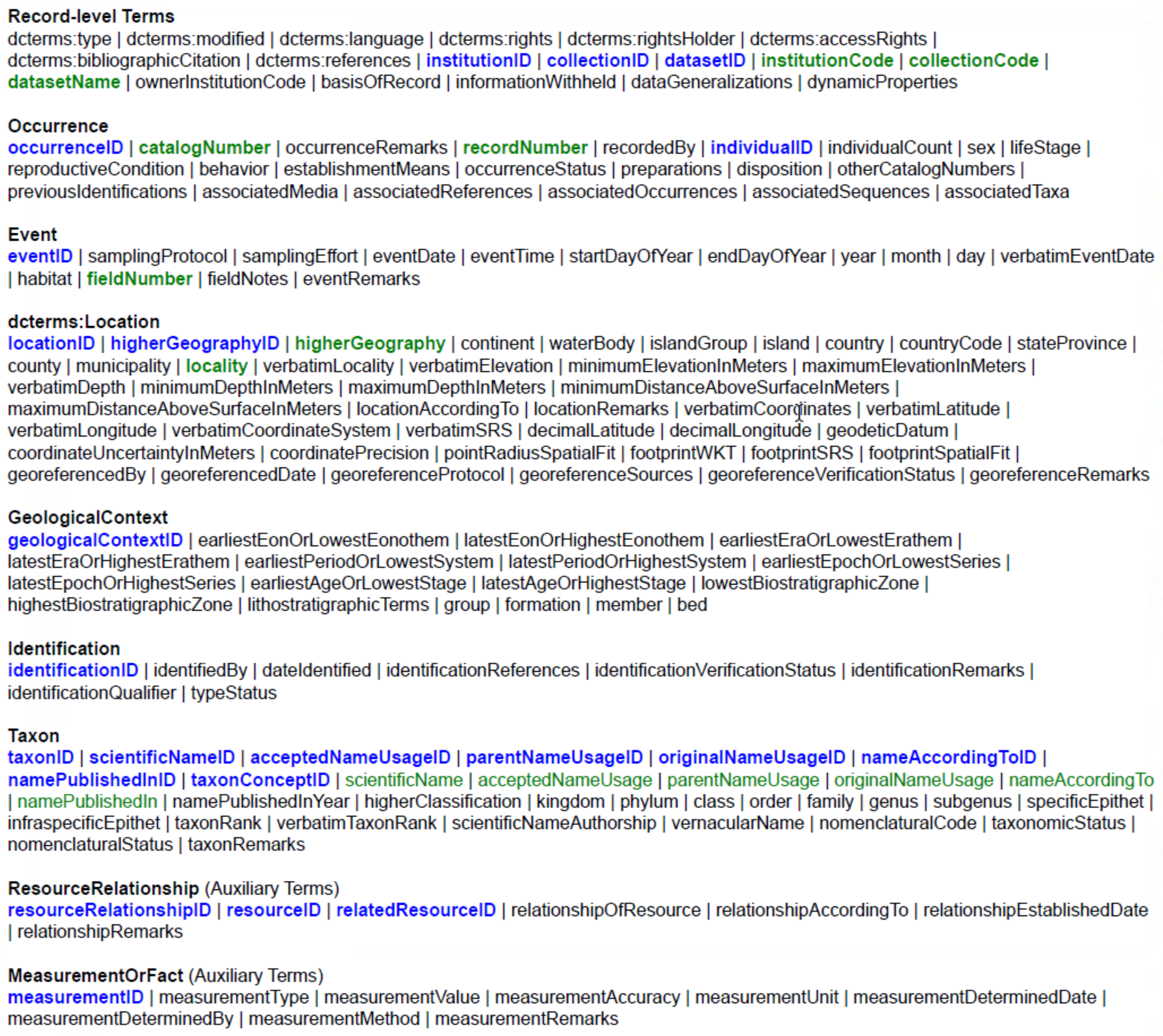
- Will vary according to Journal and each Journal offers guidance templates
- Generally the narrative includes:
    - Project details
    - Methods – Experimental or sampling design, data acquisition process, research methods
    - Coverage
    - Format
    - Quality
    - Descriptions of variables
    - License, provenance, reuse

Example of Recommended Repositories by Data Journals



Repository Name	Information on fees/costs	Size limits	Integrated with *Scientific Data*'s manuscript submission system	Re3data / BioSharing entry
Dryad Digital Repository	$120 USD for first 20 GB, and $50 USD for each additional 10 GB	None stated	**Yes ✓**	view BioSharing entry
figshare	100 GB free per *Scientific Data* manuscript. Additional fees apply for larger datasets	1 TB per dataset	**Yes ✓** - To qualify for the 100 GB of free storage, data must be uploaded to figshare via our submission system. Download instructions.	view BioSharing entry
Harvard Dataverse	Contact repository for datasets over 1 TB	2.5 GB per file, 10 GB per dataset	No	view re3data entry
Open Science Framework	None stated	5 GB per file, multiple files can be uploaded	No	view BioSharing entry
Zenodo	Cost information	50 GB per dataset	No	view re3data entry

Example of Metadata requirements - GBIF



Record-level Terms
dcterms:type | dcterms:modified | dcterms:language | dcterms:rights | dcterms:rightsHolder | dcterms:accessRights | dcterms:bibliographicCitation | dcterms:references | **institutionID** | **collectionID** | **datasetID** | **institutionCode** | **collectionCode** | **datasetName** | ownerInstitutionCode | basisOfRecord | informationWithheld | dataGeneralizations | dynamicProperties

Occurrence
occurrenceID | **catalogNumber** | occurrenceRemarks | **recordNumber** | recordedBy | **individualID** | individualCount | sex | lifeStage | reproductiveCondition | behavior | establishmentMeans | occurrenceStatus | preparations | disposition | otherCatalogNumbers | previousIdentifications | associatedMedia | associatedReferences | associatedOccurrences | associatedSequences | associatedTaxa

Event
eventID | samplingProtocol | samplingEffort | eventDate | eventTime | startDayOfYear | endDayOfYear | year | month | day | verbatimEventDate | habitat | **fieldNumber** | fieldNotes | eventRemarks

dcterms:Location
locationID | **higherGeographyID** | **higherGeography** | continent | waterBody | islandGroup | island | country | countryCode | stateProvince | county | municipality | **locality** | verbatimLocality | verbatimElevation | minimumElevationInMeters | maximumElevationInMeters | verbatimDepth | minimumDepthInMeters | maximumDepthInMeters | minimumDistanceAboveSurfaceInMeters | maximumDistanceAboveSurfaceInMeters | locationAccordingTo | locationRemarks | verbatimCoordinates | verbatimLatitude | verbatimLongitude | verbatimCoordinateSystem | verbatimSRS | decimalLatitude | decimalLongitude | geodeticDatum | coordinateUncertaintyInMeters | coordinatePrecision | pointRadiusSpatialFit | footprintWKT | footprintSRS | footprintSpatialFit | georeferencedBy | georeferencedDate | georeferenceProtocol | georeferenceSources | georeferenceVerificationStatus | georeferenceRemarks

GeologicalContext
geologicalContextID | earliestEonOrLowestEonothem | latestEonOrHighestEonothem | earliestEraOrLowestErathem | latestEraOrHighestErathem | earliestPeriodOrLowestSystem | latestPeriodOrHighestSystem | earliestEpochOrLowestSeries | latestEpochOrHighestSeries | earliestAgeOrLowestStage | latestAgeOrHighestStage | lowestBiostratigraphicZone | highestBiostratigraphicZone | lithostratigraphicTerms | group | formation | member | bed

Identification
identificationID | identifiedBy | dateIdentified | identificationReferences | identificationVerificationStatus | identificationRemarks | identificationQualifier | typeStatus

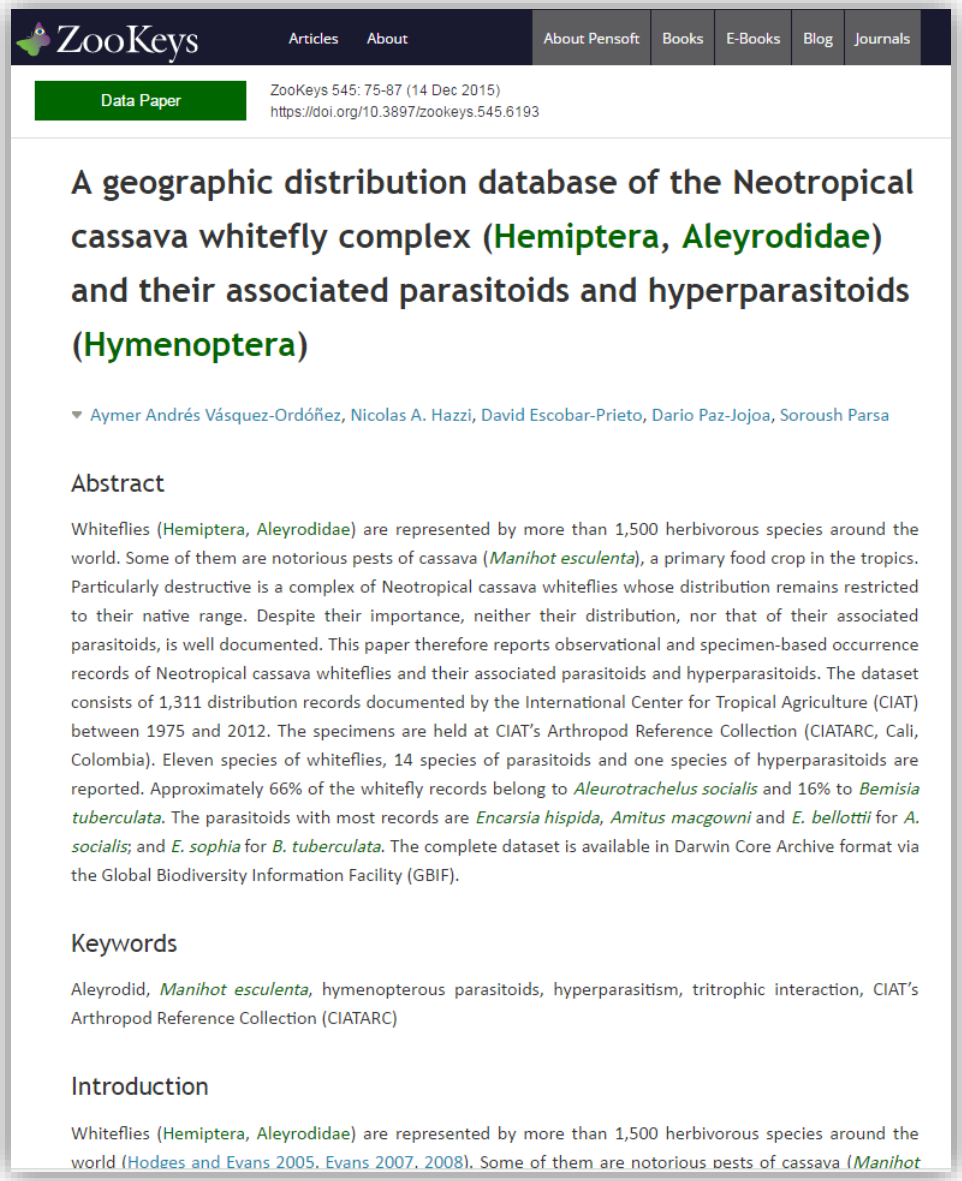
Taxon
taxonID | **scientificNameID** | **acceptedNameUsageID** | **parentNameUsageID** | **originalNameUsageID** | **nameAccordingToID** | **namePublishedInID** | **taxonConceptID** | scientificName | acceptedNameUsage | parentNameUsage | originalNameUsage | nameAccordingTo | namePublishedIn | namePublishedInYear | higherClassification | kingdom | phylum | class | order | family | genus | subgenus | specificEpithet | infraspecificEpithet | taxonRank | verbatimTaxonRank | scientificNameAuthorship | vernacularName | nomenclaturalCode | taxonomicStatus | nomenclaturalStatus | taxonRemarks

ResourceRelationship (Auxiliary Terms)
resourceRelationshipID | **resourceID** | **relatedResourceID** | relationshipOfResource | relationshipAccordingTo | relationshipEstablishedDate | relationshipRemarks

MeasurementOrFact (Auxiliary Terms)
measurementID | measurementType | measurementValue | measurementAccuracy | measurementUnit | measurementDeterminedDate | measurementDeterminedBy | measurementMethod | measurementRemarks

Credit: Daniel Amariles

Example of CIAT Data papers



ZooKeys 545: 75-87 (14 Dec 2015)
https://doi.org/10.3897/zookeys.545.6193

A geographic distribution database of the Neotropical cassava whitefly complex (Hemiptera, Aleyrodidae) and their associated parasitoids and hyperparasitoids (Hymenoptera)

Aymer Andrés Vásquez-Ordóñez, Nicolas A. Hazzi, David Escobar-Prieto, Dario Paz-Jojoa, Soroush Parsa

Abstract

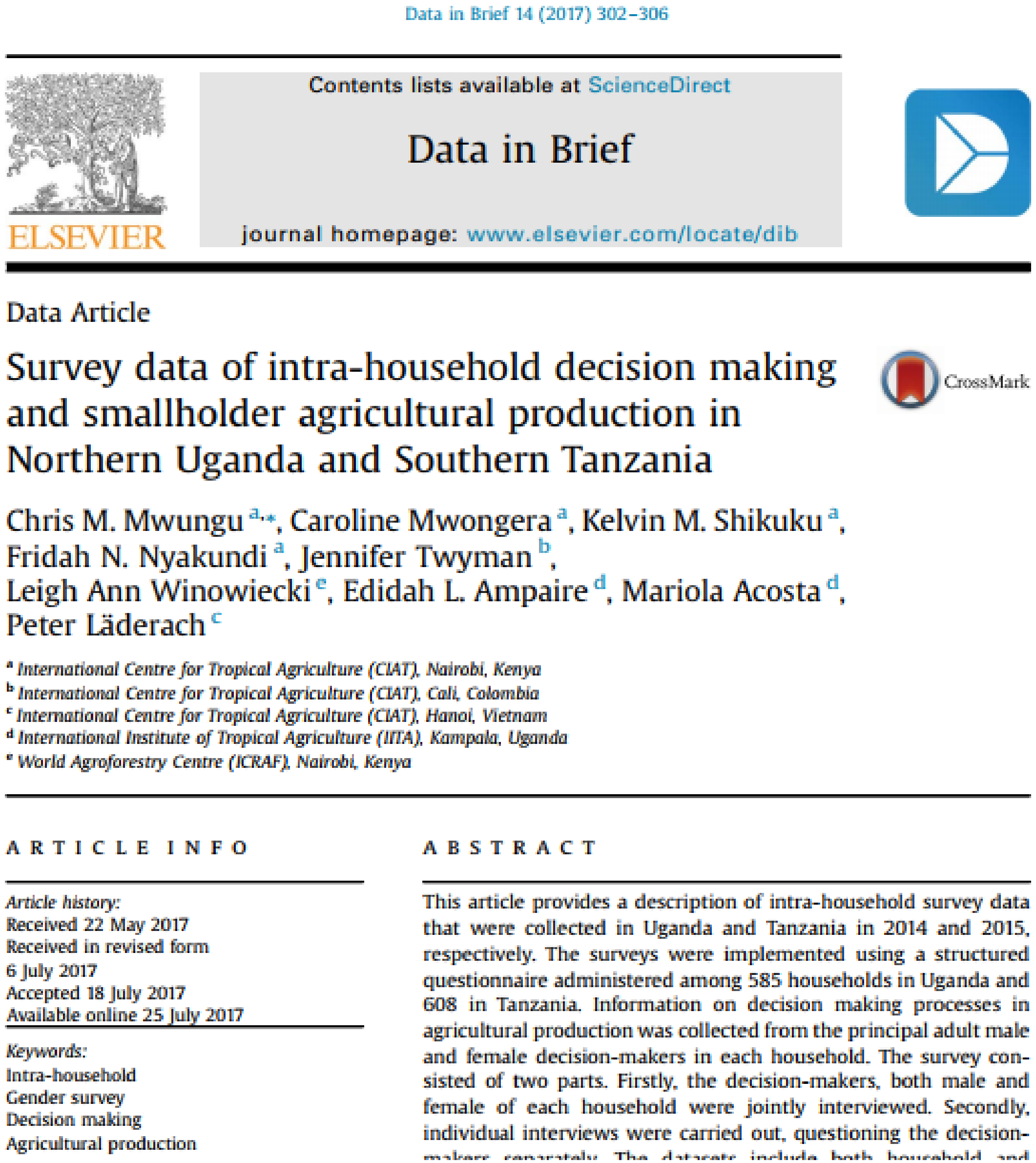
Whiteflies (Hemiptera, Aleyrodidae) are represented by more than 1,500 herbivorous species around the world. Some of them are notorious pests of cassava (*Manihot esculenta*), a primary food crop in the tropics. Particularly destructive is a complex of Neotropical cassava whiteflies whose distribution remains restricted to their native range. Despite their importance, neither their distribution, nor that of their associated parasitoids, is well documented. This paper therefore reports observational and specimen-based occurrence records of Neotropical cassava whiteflies and their associated parasitoids and hyperparasitoids. The dataset consists of 1,311 distribution records documented by the International Center for Tropical Agriculture (CIAT) between 1975 and 2012. The specimens are held at CIAT's Arthropod Reference Collection (CIATARC, Cali, Colombia). Eleven species of whiteflies, 14 species of parasitoids and one species of hyperparasitoids are reported. Approximately 66% of the whitefly records belong to *Aleurotrachelus socialis* and 16% to *Bemisia tuberculata*. The parasitoids with most records are *Encarsia hispida*, *Amitus macgowni* and *E. bellottii* for *A. socialis*; and *E. sophia* for *B. tuberculata*. The complete dataset is available in Darwin Core Archive format via the Global Biodiversity Information Facility (GBIF).

Keywords

Aleyrodid, *Manihot esculenta*, hymenopterous parasitoids, hyperparasitism, tritrophic interaction, CIAT's Arthropod Reference Collection (CIATARC)

Introduction

Whiteflies (Hemiptera, Aleyrodidae) are represented by more than 1,500 herbivorous species around the world (Hodges and Evans 2005, Evans 2007, 2008). Some of them are notorious pests of cassava (*Manihot*



Data in Brief 14 (2017) 302–306

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Survey data of intra-household decision making and smallholder agricultural production in Northern Uganda and Southern Tanzania

Chris M. Mwungu [a,*], Caroline Mwongera [a], Kelvin M. Shikuku [a], Fridah N. Nyakundi [a], Jennifer Twyman [b], Leigh Ann Winowiecki [e], Edidah L. Ampaire [d], Mariola Acosta [d], Peter Läderach [c]

[a] International Centre for Tropical Agriculture (CIAT), Nairobi, Kenya
[b] International Centre for Tropical Agriculture (CIAT), Cali, Colombia
[c] International Centre for Tropical Agriculture (CIAT), Hanoi, Vietnam
[d] International Institute of Tropical Agriculture (IITA), Kampala, Uganda
[e] World Agroforestry Centre (ICRAF), Nairobi, Kenya

ARTICLE INFO

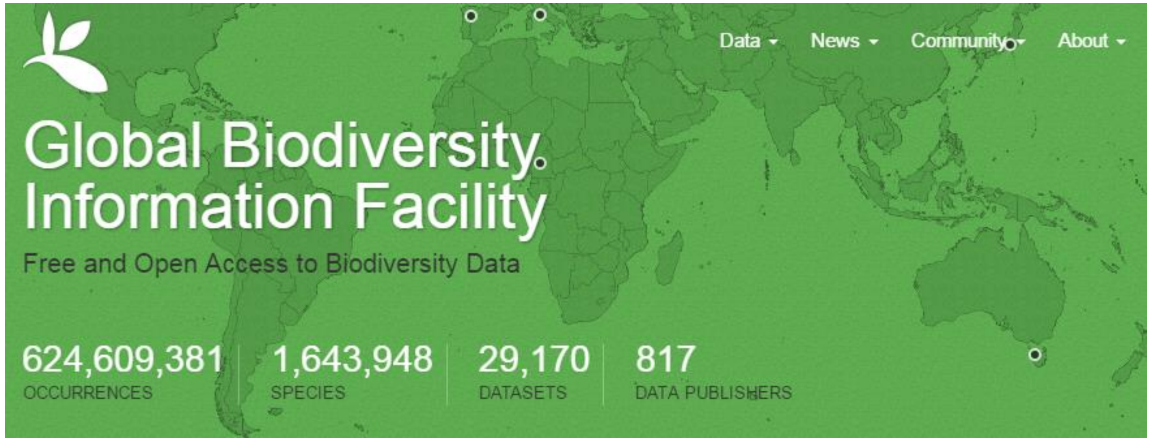
Article history:
Received 22 May 2017
Received in revised form
6 July 2017
Accepted 18 July 2017
Available online 25 July 2017

Keywords:
Intra-household
Gender survey
Decision making
Agricultural production

ABSTRACT

This article provides a description of intra-household survey data that were collected in Uganda and Tanzania in 2014 and 2015, respectively. The surveys were implemented using a structured questionnaire administered among 585 households in Uganda and 608 in Tanzania. Information on decision making processes in agricultural production was collected from the principal adult male and female decision-makers in each household. The survey consisted of two parts. Firstly, the decision-makers, both male and female of each household were jointly interviewed. Secondly, individual interviews were carried out, questioning the decision-makers separately. The datasets include both household and

Where: Subject Specific Repositories



Global Biodiversity Information Facility

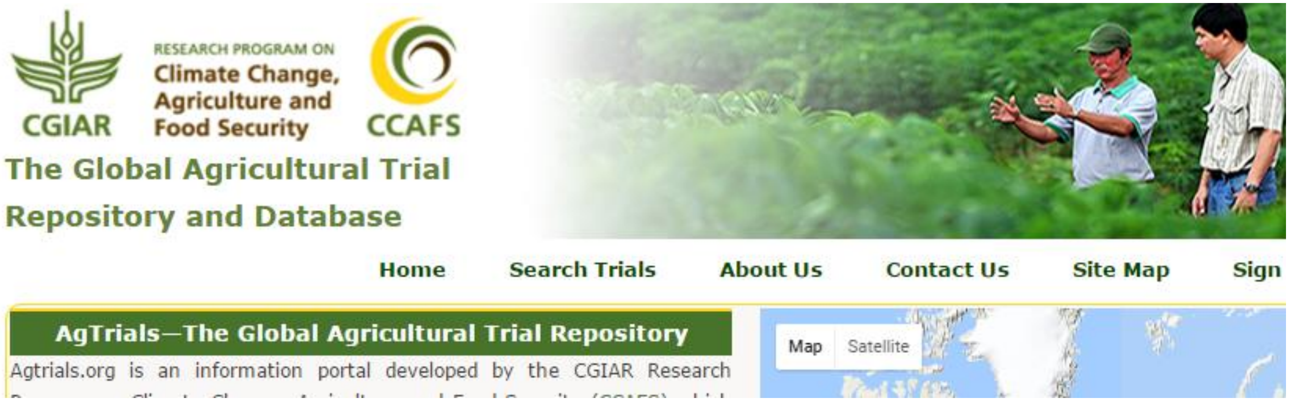
Free and Open Access to Biodiversity Data

624,609,381 OCCURRENCES

1,643,948 SPECIES

29,170 DATASETS

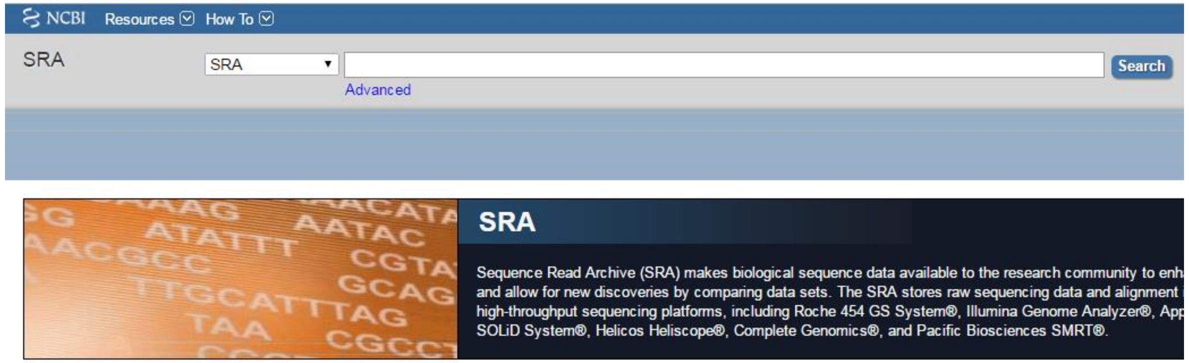
817 DATA PUBLISHERS



The Global Agricultural Trial Repository and Database

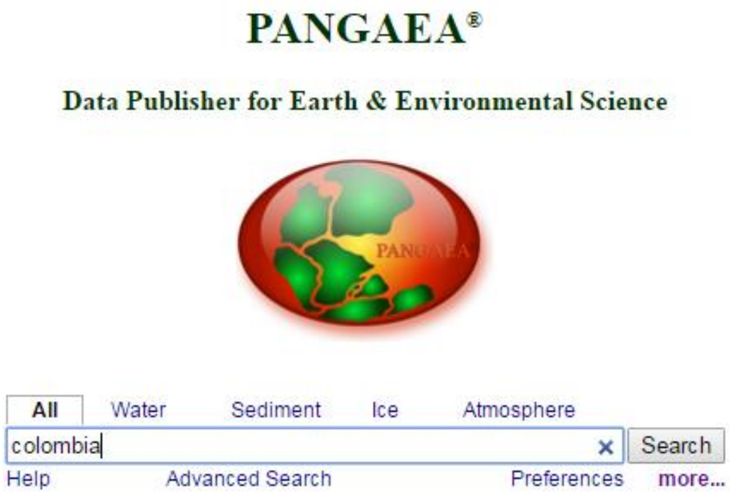
AgTrials—The Global Agricultural Trial Repository

Agtrials.org is an information portal developed by the CGIAR Research



SRA

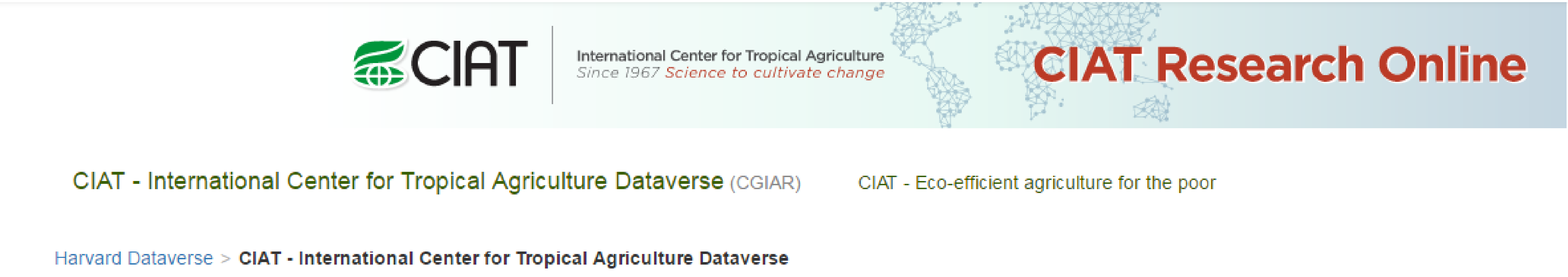
Sequence Read Archive (SRA) makes biological sequence data available to the research community to enh
and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment
high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, App
SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.



PANGAEA®

Data Publisher for Earth & Environmental Science

Where: Institutional Repositories

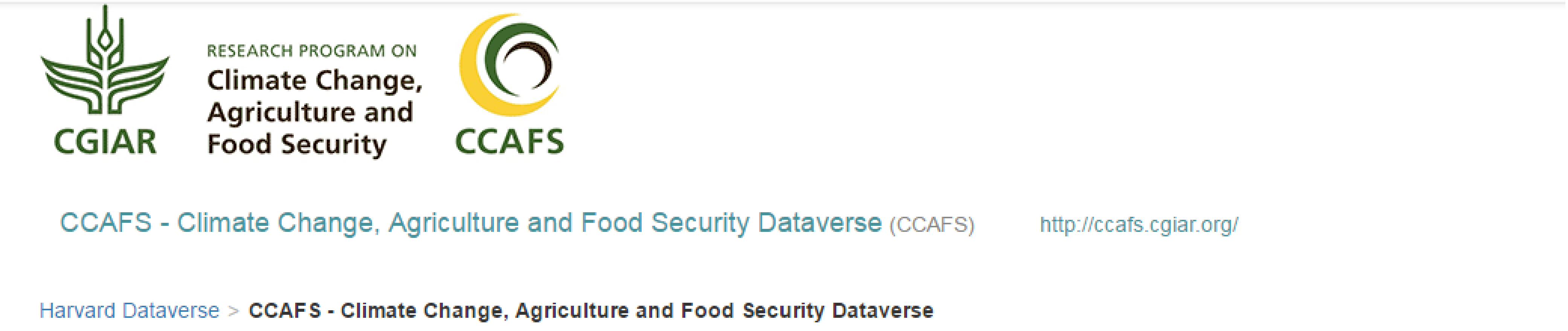


International Center for Tropical Agriculture
Since 1967 Science to cultivate change

CIAT Research Online

CIAT - International Center for Tropical Agriculture Dataverse (CGIAR)   CIAT - Eco-efficient agriculture for the poor

Harvard Dataverse > CIAT - International Center for Tropical Agriculture Dataverse



RESEARCH PROGRAM ON
Climate Change,
Agriculture and
Food Security

CGIAR   CCAFS

CCAFS - Climate Change, Agriculture and Food Security Dataverse (CCAFS)   http://ccafs.cgiar.org/

Harvard Dataverse > CCAFS - Climate Change, Agriculture and Food Security Dataverse

Where: General Repositories

figshare
credit for all your research

DRYAD

zenodo

MENDELEY DATA

Where: supplementary material to your research paper

- May be used for smaller datasets
- Ensure you are not transferring copyright of the data to publisher
- Often not all the primary data that underlies a publication

Example of a publishing Workflow

1	Publish peer reviewed data paper	Submit paper for review with references & DOI to data	Paper published and references published data paper
2	Submit data in CIAT Dataverse and do not publish	Submit paper for review Submit data for review via private link	Paper published Data published by CIAT Paper reference dataset
3	Submit paper for review	Paper approved and published. Submit data to CIAT Dataverse	Data published on CIAT Dataverse and references paper.

Restrictions to Sharing Data

Privacy – Information that identifies and individual

Confidentiality – Information that should not be shared

Security – Release of data will cause threats to someone or something

Data Licenses

For your published data to be truly open, reusable and redistributable you need to apply a license that guarantees the openness of the data.

Examples of open licenses applicable to data

- Creative commons Attribution 4.0 – CC-BY
- Open Data Commons Attribution – ODC–BY
- Creative Commons Public Domain – CC0
- Open Data Commons Public Domain Dedication – PDDL

When to Publish

According to CGIAR open access and open data policy.

- Data and datasets should be published within **12 months** of an appropriate project milestone such as, the end of data collection or the end of the project.

- For datasets used in publications these should be published within **6 months** of article publication.

FAQ

Will my published data (with citation and DOI) be considered “Prior publication” by the Journal I want to publish?

- Need to verify with the journal you are targeting. Many like journals from Nature, Science, Elsevier, PLoS, SAGE, BMC allow work based on prior published datasets. Others may not.

Summary

- Data is a first class research product
- Prepare your data and documentation before hand
- Choose your publication workflow
- Publish in legitimate data journals or approved repositories
- Apply a data license

International Data Week - Data Sprint

Awardee	Award
1. CIAT Data author with the most competition points.	Funding to cover the cost of participation in 1 scientific conference and article processing costs (APCs) for 1 open access journal article (total funding up to **USD 5000**).
2. CIAT Research Area with the most competition points	Funding to cover the costs of 3 open access journal articles (up **to USD 7000**) as decided by the Research Area Director.
3. Data authors with most competition points in each of CIAT regions. (The overall competition winner will be ineligible for this award)	Funding to cover the cost of 1 open access journal article each (up to USD **1000 each**).