Développer les entrepôts de données de recherche en Afrique de l'Ouest

Alex de Sherbinin, PhD
NASA Socioeconomic Data and Applications Center (SEDAC)
Center for International Earth Science Information Network (CIESIN)
Columbia University

Vice President, Conseil Technique du Systeme Mondiale des Donnees (WDS)

Science Ouverte dans le Sud

23 octobre 2019
Dakar, Senegal

Vue d'ensemble

- Qu'est-ce que le Système Mondial des Donnees (WDS) et pourquoi devenir membre?
- Tendences en science et donnees ouvertes
- Introduction à la certification et CoreTrustSeal
- Expliquer le processus d'accréditation pour devenir membre de WDS
- Processus de gestion et dissémination des données: exemple du NASA SEDAC
- Questions pour discussion

ICSU
WORLD DATA SYSTEM

Apercu du WDS

ICSU
WORLD DATA SYSTEM

Qu’est ce que WDS?

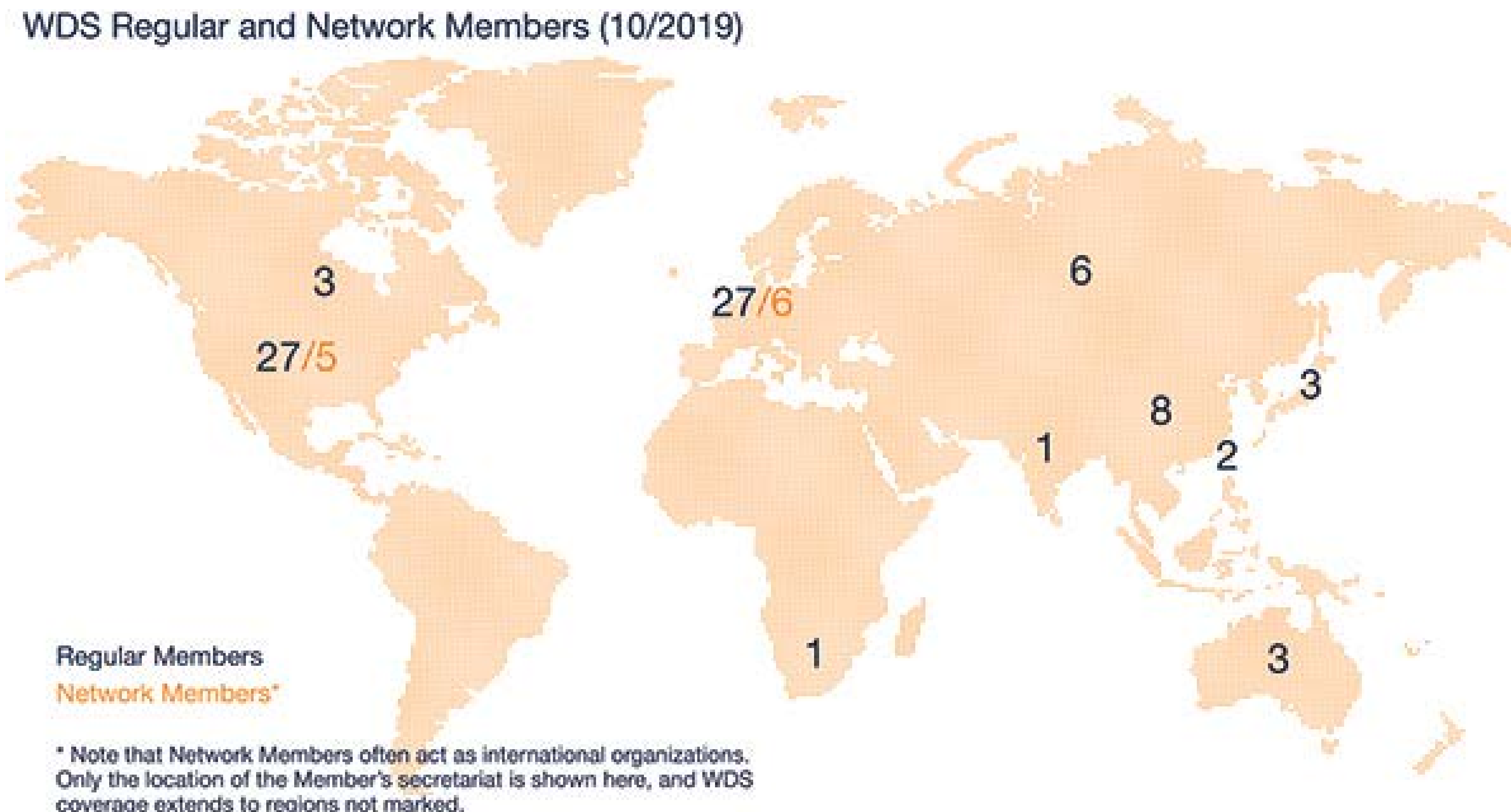
Le World Data System a pour mission de promouvoir une gestion **à long terme et un accès universel et équitable** à des données scientifiques de qualité garantie ainsi qu'à des **services, produits et informations de données de qualité**, dans diverses disciplines des sciences naturelles et sociales.

Pour ce faire, WDS coordonne et soutient des **services de données scientifiques fiables pour la fourniture, l'utilisation et la conservation d'ensembles de données pertinents** afin de faciliter la recherche scientifique dans le cadre du CIS, tout en renforçant leurs liens avec le monde de la recherche.

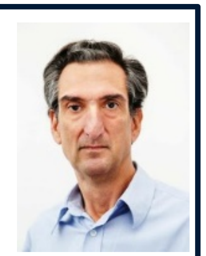
Il y a actuellement 81 membres ordinaires, 11 membres du réseau, 11 membres partenaires et 20 membres associés.

ICSU WORLD DATA SYSTEM

Distribution geographique de membres



Membre du comite scientifique : Alfredo Tolmasquim, Scientific Director of the Museum of Tomorrow, and Museum of Astronomy and Related Sciences, History of Science Coordination, Rio de Janeiro, Brazil



Membre recente du comite scientifique: Research Director at Institute of Research for Development (IRD) and Laboratoire d’Etude des Transferts en Hydrologie et Environnement (LTHE), University of Grenoble-Alpes, France. Laboratoire de Physique de l'Atmophère et Mécanique des fluides, Université Félix Houphouet Boigny, Abidjan, Côte d'Ivoire

ICSU
WORLD DATA SYSTEM

Pourquoi joigner?

Réputation améliorée avec l'imprimatur du CIS

Une visibilité accrue dans les activités internationales améliore la réputation et la base d'utilisateurs

Amélioration des perspectives de financement

Interactions et échange de données avec d'autres membres

Démontrez l'engagement de votre centre de données en faveur de la science ouverte

Être considéré comme un service de données scientifiques fiable

Améliorez vos pratiques et processus

Membre Candidate

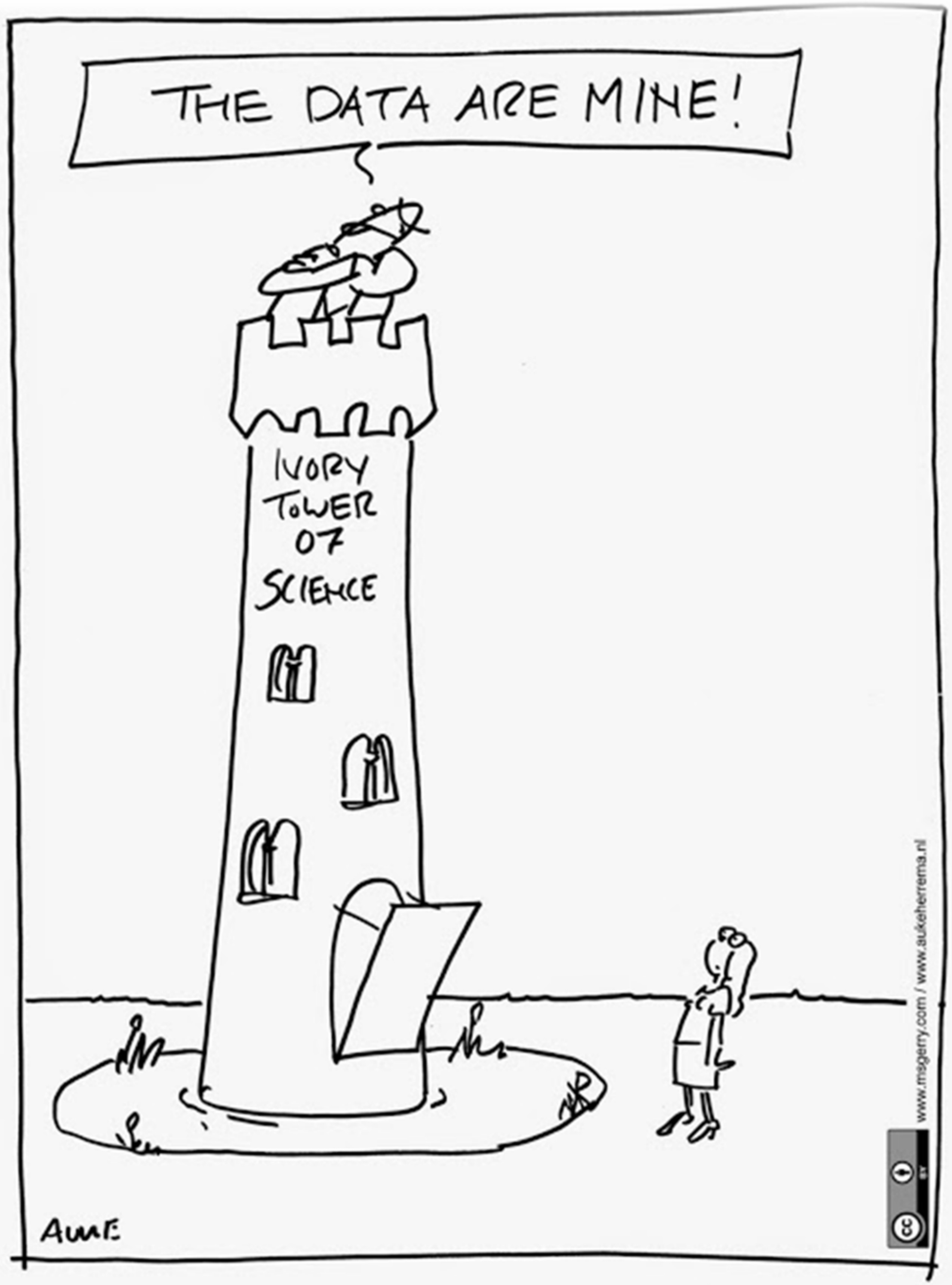
Une nouvelle catégorie de membre disponible cette année. Suivre ces étapes:

1. Envoyez une expression d'interet pour etre une membre reguliere
2. Ajoutez dans les commentaires qu'il vous faute un peu plus de temps pour compléter les étapes de certification
3. Utiliser le formulaire "CTS" et répondez de meilleur façon possible

Le WDS va vous guider et soutenir sur le processus a travers le bureau du programme (IPO), bureau de technologie (ITO) et conseil scientifique

ICSU
WORLD DATA SYSTEM

Tendences en science et donnees ouvertes



SCENE FROM THE PAST?

Les données sont les miens!

ICSU
WORLD DATA SYSTEM

Deux modèles de politiques

L'ancienne regime

- Information c'est le pouvoir
- Il faut payer pour les données
- Ceux qui produisent les données doivent récupérer les coutes de production
- Les utilisateurs ne peuvent pas faire "value add" aux produits a cause des licences restrictives

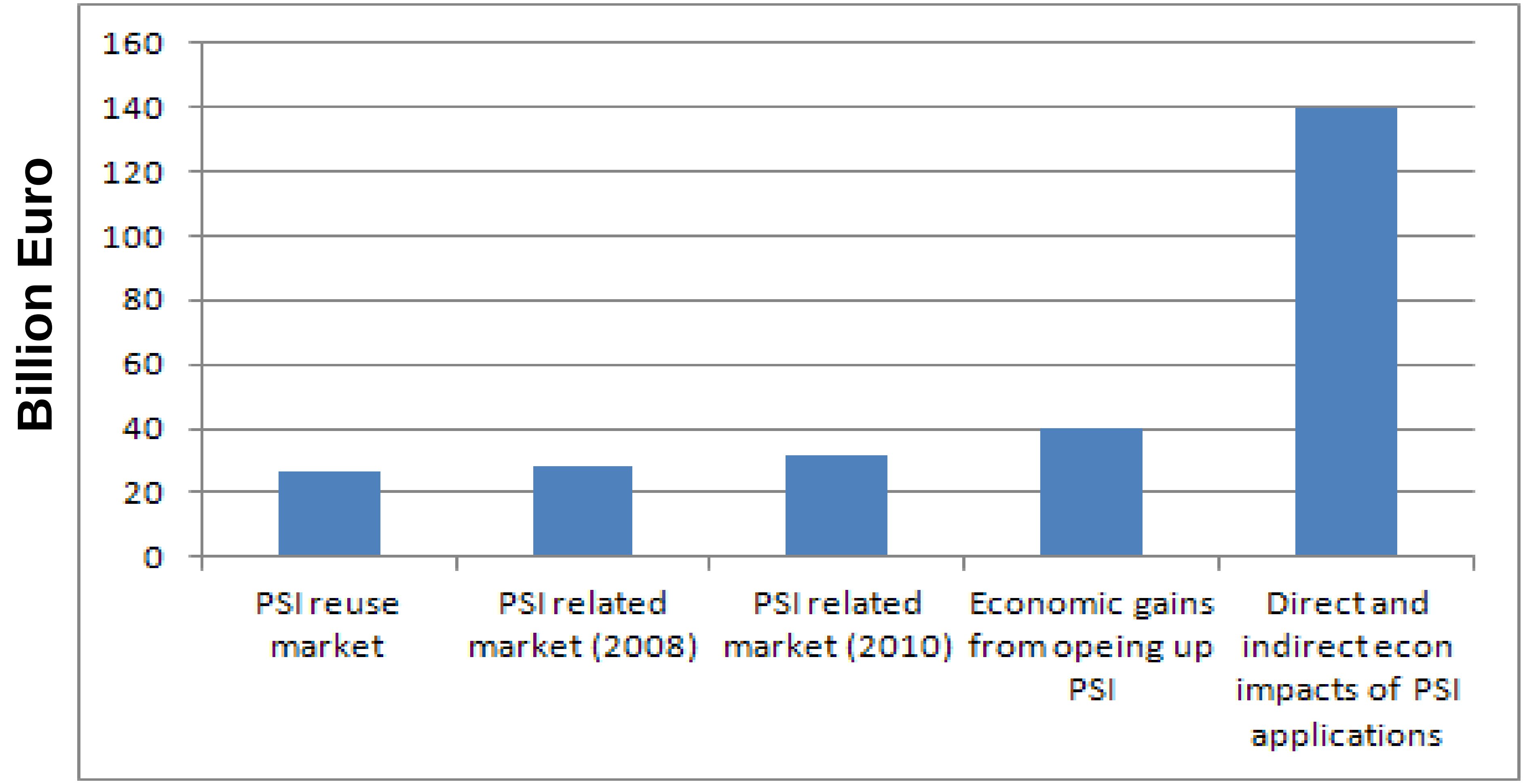
Nouvelle regime

- Access aux données gratuites et sans contraintes
- Une société mieux informée
- Politique publique informée par l'information scientifique
- Accélération de l'innovation
- Les coutes a l'industrie baisse
- Secteur d'information grandisse
- Les impots sur le secteur subvention developpement des donnees
- Allez des <u>données</u> aux <u>services</u>

ICSU WORLD DATA SYSTEM

Investissement dans les données ouvertes

Economic benefits of open public sector information in the EU27



Source: Vickery (2011), "Review of Recent Studies on PSI Re-Use and Related market Developments"

ICSU WORLD DATA SYSTEM

Global Data Sharing Trends*

Over 4,600 Wiley authors from 112 countries completed our 2016 Wiley Open Science Researcher Insights Survey

By collating results of our Wiley authors from surveys on Open Science topics in 2013, 2014, and 2016, we have started to build a valuable dataset for analysis and trend identification. Despite geographical and subject-level differences among authors, there are underlying commonalities in Open Science practices. The insights reported by our respondents show a willingness to move forward with open initiatives, but confusion around the best ways to do so.

Data accessibility trends

- **56%** Spent a large amount of time to make their data reproducible
- **43%** Used other researchers' publicly available data
- **47%** Checked another paper's source data

Data sharing in 2016

31% do not share data

69% share data

More than two thirds of Wiley researchers reported they are now sharing their data. Though this varies geographically and across research disciplines we are seeing that more researchers are sharing their data and taking efforts to make it reproducible. Archiving in institutional repositories, public repositories, and personal web pages has almost doubled since 2014.

Top 4 researcher motivations for sharing data

- **39%** Increase the impact and visibility of my research
- **35%** Public benefit
- **31%** Transparency and re-use
- **29%** Journal requirement

Top 4 reasons why researchers are hesitant to share their data

1. **50%** - Intellectual property or confidentiality issues
2. **31%** - Ethical concerns
3. **23%** - I am concerned about misinterpretation or misuse of my research
4. **22%** - I am concerned that my research will be scooped

Ways data is shared

- **41%** As supplementary material in a journal
- **10%** Discipline-specific data repository (e.g. GenBank, OpenEI, Protein Data Bank, TreeBASE)
- **29%** Personal, institutional, or project webpage
- **6%** General-purpose data repository (e.g. Dryad, figshare)
- **25%** Institutional data repository (i.e. university or institute-sponsored)

Researchers also report sharing their data in other ways including: 49% are sharing their data at conferences while 34% of researchers share their data upon informal request (email, direct contact, etc).

Researchers sharing data by region

*Sharing data includes data the researchers have produced and shared.

WILEY

Principes des données ouvertes du WDS (1)

- **Data, metadata, products, and information should be fully and openly shared**, subject to national or international jurisdictional laws and policies, including respecting appropriate extant restrictions, and in accordance with international standards of ethical research conduct.
- Data, metadata, products, and information produced for research, education, and public-domain use **will be made available with minimum time delay and free of charge**, or for **no more than the cost of dissemination**, which may be waived for lower-income user communities to support equity in access.

Principes des données ouvertes du WDS (2)

- All who produce, share, and use data and metadata are stewards of those data, and have responsibility for ensuring that the authenticity, quality, and integrity of the data are preserved, and respect for the data source is maintained by ensuring privacy where appropriate, and encouraging appropriate citation of the dataset and original work and acknowledgement of the data repository.
- **Data should be labelled 'sensitive' or 'restricted' only with appropriate justification and following clearly defined protocols**, and should in any event be made available for use on the least restrictive basis possible.

ICSU WORLD DATA SYSTEM

FULL TEXT ARTICLE

Closing the door on parachutes and parasites

The Lancet Global Health

Lancet Global Health, 2018-06-01, Volume 6, Issue 6, Pages e593-e593,

No one likes a parachute researcher: the one who drops into a countr
infrastructure, personnel, and patients, and then goes home and wr
journal. At *The Lancet Global Health* , we look extremely unfavou
who have done primary research in another country (particularly
but not included any author from that nation. For research invo
existing facilities, and follow-up, the notion that no locally bas
contribution" (per authorship criteria) to the acquisition of da
out, none of those individuals additionally fulfilled the criter
the study or writing of the report, but then perhaps they we
had, the design would have been more appropriate to the
realistic.

FAIRER COLLABORATION

We need to end "parachute" research which sidelines the work of African scientists

By Moses John Bockarie/Njala University • January 29, 2019

Sign up for the Quartz Daily Brief email

Enter your email

Sign me up

Stay updated about Quartz products and events.

There's growing condemnation of "parachute research" among the global scientific community. This refers to the practice of scientists and research groups from the global north conducting research and collecting data in poorer parts of the world, publishing their findings in prestigious journals – and giving little or no credit to their local collaborators.

Points of view expressed in recent CODATA discussion on “digital colonialism” (Oct. 2019)

- “A few years ago I was working with a young African researcher on an agro-forestry research project. No sooner had we started than I realised that her team had only some descriptive statistics but no direct access to the biomass data which she and her colleagues had spent months collecting from two islands! The vast chunk of the data had left with the development partners at the end of the project. It turned out, nobody at the centre had any knowledge or pressing interest to pursue the data and there was already new initiatives to run another project, which in was almost a duplicate of the first, but this time with a different development partner.” …. “Apparently, what CODATA, WDS, RDA and many others are doing is to enhance global utilisation of data in human development. **The key question is why, in the 21st century (the Big Data era), issues like the ones we have been discussing over the last 24 hours are still common place? The answer lies in the imbalance in the ownership of data resources - tools for acquisition, storage, analysis and dissemination**.” – Kassim Mwitondi, Sheffield Hallam University
- “I also recently learned that **data from NGOs are at best perhaps shared with some ministries in Bamako, Mali, but not within a region in which the work is being done. This implies that local decision makers remain dependent on the information/data stream back from the ministries which may take some months, if ever**. This can negate the purpose of the work executed.” - Niek van Duivenbooden, Trimpact

Points of view expressed in recent CODATA discussion on “digital colonialism” (Oct. 2019)

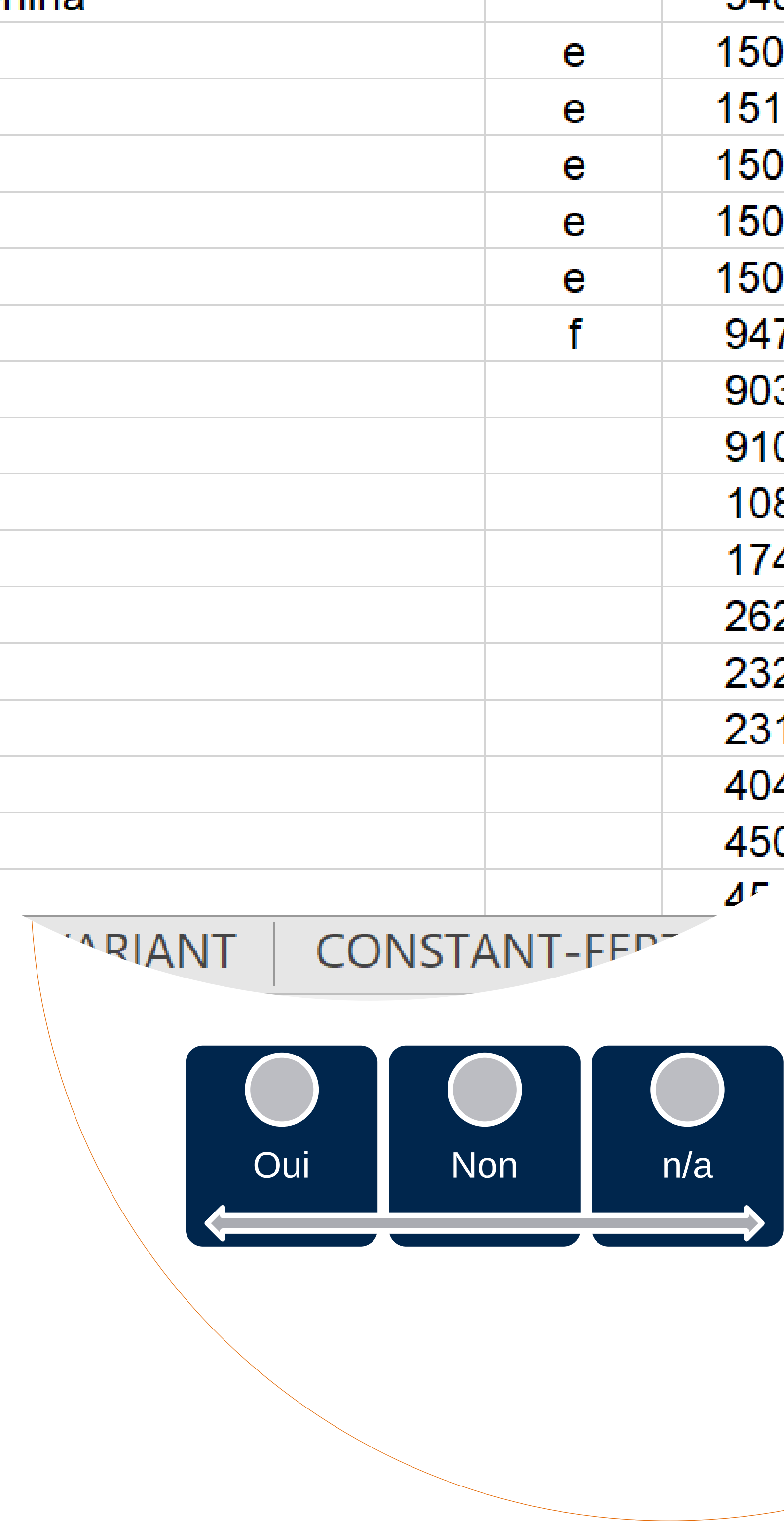
- “This is not a new problem and **there are too many examples of well meaning projects financed from the outside which rarely continue after the project has been completed.** Even in the best circumstances when adequate descriptions of the project and associated data and metadata are documented, which as Ernie and others have pointed out is too rarely the case, the issue of effective continuity continues to be a challenge. **My own experience suggests that a key element in improving the situation is meaningful national involvement and ownership of such projects from the outset including project formulation**.” – Fraser Taylor, Carleton University, Canada
- “Together with several colleagues (Kenya, Botswana) we have been doing some work for the **African Science Granting Councils (19 African States) that analyses advantages and disadvantages for Africa of federated open science practices, together with required to deliver them to best effect**. The issues you have all described are being addressed, such that we hope the Granting Councils will address them, together with International Partners helpful if we were able to call on your experiences as evidence.” – Geoffrey Boulton, CODATA

Solutions expressed in recent CODATA discussion on “digital colonialism” (Oct. 2019)

- “The GODAN Africa Agenda is to promote a consensus on an Open data paradigm that anonymizes data for access, use and reuse for innovation. Intra-country data needs to be as raw as the users may seek to do. **It makes sense for each country to plan its own strategy on hosting, but GODAN supports countries to think through the internal processes. Alone, however, we are not able to do it which is why we are a network to work with those who share our vision of making evidence available for decision making using data openness**. The best locale for the kind of discussion we seem to hold is at the AU or ECA, who could kick start the discussions then regions and countries can take that forward but from the perspective of breaking the big pan African challenge into smaller achievable milestones.” – Kiringai Kamau, GODAN

Le défi

- Pour prévenir l’expatriation des donnes africaines if faut créer les entrepôts de données nationaux et régionaux
- Il faut créer la capacité de gérer des données d’une manière fiable a longue terme
- Ca ne veut pas dire qu’on ne peut pas utiliser les services “cloud” outre de l'Afrique … mais que les priorités et les responsabilités de la préservation restent dans les mains des institutions africains



ICSU
WORLD DATA SYSTEM

Introduction à la certification et CoreTrustSeal

https://www.rd-alliance.org/coretrustseal-criteres-de-conformite

https://www.ouvrirlascience.fr/entrepots-de-donnees-de-confiance-criteres-de-conformite/

Piliers de la confiance

- Les entrepôts de données doivent adhérer aux principes de « TRUST » → **T**ransparence, **R**esponsabilité, communauté d'**U**tilisateurs, **d**urabilité et **T**echnologie

- Réputation de l'entrepot (*intégrité, transparence, compétence, prévisibilité, garanties, intentions positives*)

- *la reconnaissance externe:*
  - *réputation (chercheurs)*
  - *l'approbation des bailleurs de fonds, éditeurs*

"Perhaps the biggest challenge in sharing data is trust: how do you create a system robust enough for scientists to trust that, if they share, their data won't be lost, garbled, stolen or misused?"

Pourquoi une certification formelle?

Assurer que le centre est « de confiance »

Mais… il a peut-être déjà la confiance de ses utilisateurs…

L'exemple du Centre de Données astronomiques de Strasbourg (CDS)

- Crée en 1972
- Centre de données de référence pour la communauté astronomique internationale
- Infrastructure de Recherche sur la Feuille de Route nationale
- ~1 000 000 requêtes/jour sur les services

ICSU
WORLD DATA SYSTEM

Source: Francoise Genova

Oui, pourquoi?

Critères établis par des personnes compétentes et applicables quel que soit le cadre disciplinaire

Evaluation externe par des personnes compétentes

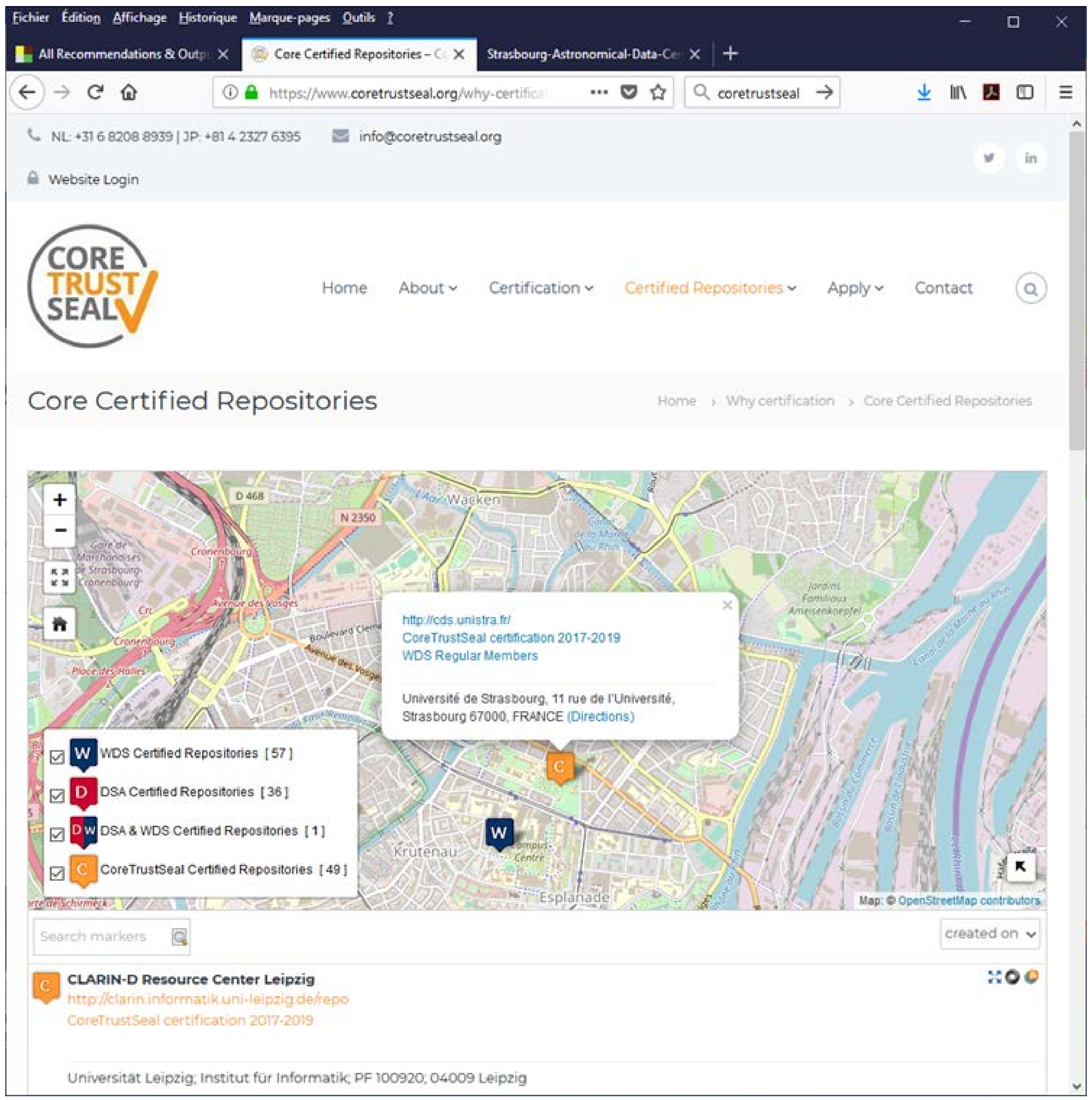
Au préalable, auto-évaluation selon les critères, qui permet de vérifier l’organisation et les processus et d’identifier des améliorations possibles

Un point important dans les Data Management Plans

Source: Francoise Genova

ICSU WORLD DATA SYSTEM

Le CDS est certifié



World Data System - WDS

Data Seal of Approval - DSA

CoreTrustSeal - CTS

Document produit pour la certification CoreTrustSeal:

https://www.coretrustseal.org/wp-content/uploads/2019/02/Strasbourg-Astronomical-Data-Centre.pdf

Source: Francoise Genova

Les critères du CoreTrustSeal

Source: Francoise Genova

La certification CTS

Toute l'information est sur le site de CoreTrustSeal

– https://www.coretrustseal.org/

Contexte + 16 critères

Document pour guider les évaluateurs et les candidats

– En cours, V1.1 2017-2019

https://www.coretrustseal.org/wp-content/uploads/2017/01/20180629-CTS-Extended-Guidance-v1.1.pdf

– Traduction française par RDA France

https://www.rd-alliance.org/coretrustseal-criteres-de-conformite

Guide en cours de révision pour 2020-2022 - Les critères ne changent pas!

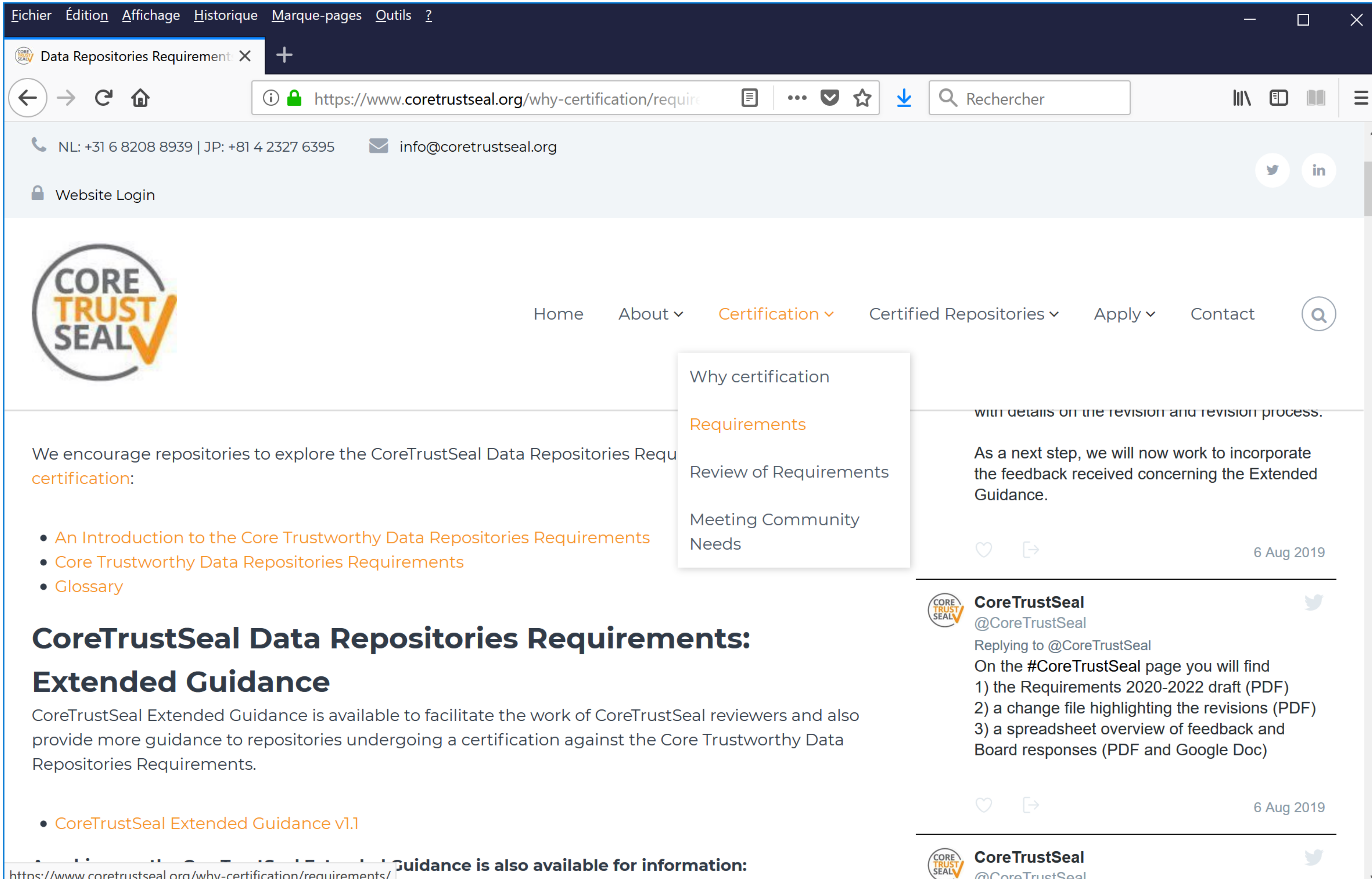
– Version préliminaire et modifications

https://www.coretrustseal.org/why-certification/review-of-requirements/

« Administrative fee » 1000€ pour 3 ans

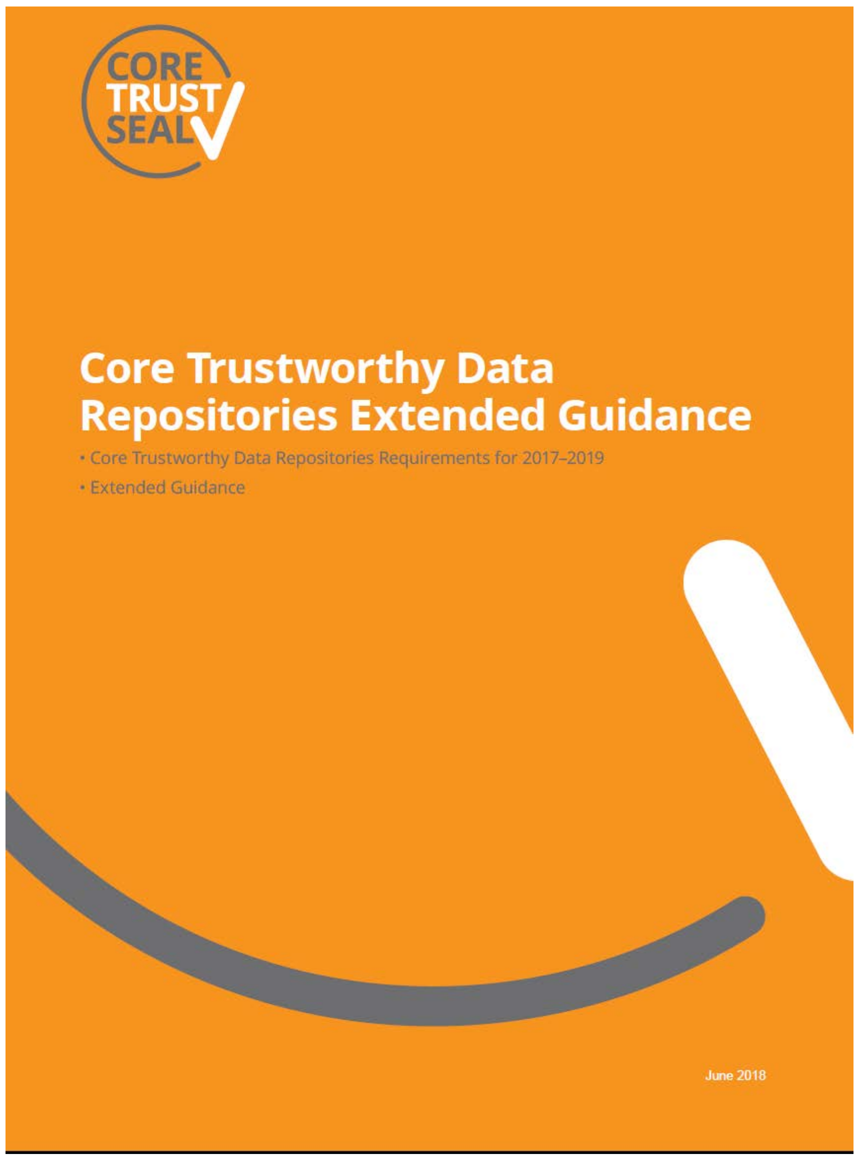
Source: Francoise Genova

Les critères sur le site CTS



Source: Francoise Genova

Les critères de certification CTS



Le contexte

16 critères, 3 thèmes:

- Infrastructure organisationnelle
- Gestion des objets numériques (données et des metadonnées)
- Technologie

Critères + aide

https://www.coretrustseal.org/why-certification/requirements/

ICSU
WORLD DATA SYSTEM

Source: Francoise Genova

11-12/09/2019

Le contexte

Type d’entrepôt

Brève description de l’entrepôt

Brève description de la communauté concernée

Niveau de curation

- Contenu en accès tel que déposé
- Curation de base (p. ex. vérification rapide, ajout de métadonnées de base ou de documentation)
- Curation avancée (p. ex. conversion vers de nouveaux formats, amélioration de la qualité de la documentation)
- Curation au niveau des données

Partenaires

Résumé des modifications depuis la candidature précédente (s’il y a lieu)

Autres informations pertinentes

Source: Francoise Genova

Infrastructure organisationnelle

R1 – Mission/périmètre

R2 – Licenses

R3 – Continuité de l’accès

R4 – Confidentialité/éthique

R5 – Infrastructure organisationnelle

R6 – Conseils d’experts

ICSU
WORLD DATA SYSTEM

Source: Francoise Genova

Gestion des objets numériques

R7 – Intégrité et authenticité des données

R8 – Appréciation et sélection des données

R9 – Procédures d’archivage documentées

R10 – Plan de préservation

R11 – Qualité des données

R12 – Processus de traitement (Workflows)

R13 – Découverte et identification des données

R14 – Réutilisation des données

ICSU
WORLD DATA SYSTEM

Source: Francoise Genova

Technologie

R15 – Infrastructure technique

R16 – Sécurité

Source: Francoise Genova

Coute et bénéfices de la certification

Quelques semaines de travail d’équipe (tout compris)

Evaluation interne

Evaluation externe

Importance croissante pour les financeurs des centres de données et des projets (DMP)

Source: Francoise Genova

ICSU WORLD DATA SYSTEM


CORE
TRUST
SEAL


Welcome.
We are building an inclusive community.

Processus de gestion et dissémination des données: exemple du NASA SEDAC

Center for International Earth Science Information Network
EARTH INSTITUTE | COLUMBIA UNIVERSITY

ICSU WORLD DATA SYSTEM

CIESIN vue d'ensemble

- Centre de l'Institute de la Terre (the Earth Institute) de Columbia University depuis 1998
- Emphase sur les données spatiales, analyses et recherche, et gestion des données
- 45 personnel des sciences sociales, naturelles, technologie, et gestion des données en 3 divisions:
  - Applications Scientifiques
  - Applications Geospatiales
  - Technologie

- CIESIN est prééminent en matière de données et applications spatiales, et s'engage avec ISC, GEO, OGC, et d'autres
- Nous avons financement de:
  - NASA (SEDAC, ROSES, GEO)
  - USAID (SERVIR, WA BiCC, etc.)
  - Bill and Melinda Gates Foundation
  - Banque Mondiale
  - Facebook
  - et d'autres…

NASA Socioeconomic Data and Applications Center (SEDAC)

La mission de SEDAC est de développer et opérer les applications qui soutien l’integration des données socioéconomiques and de la télédétection et de combler la lacune entre les sciences sociales et de la terre.

http://sedac.ciesin.columbia.edu/

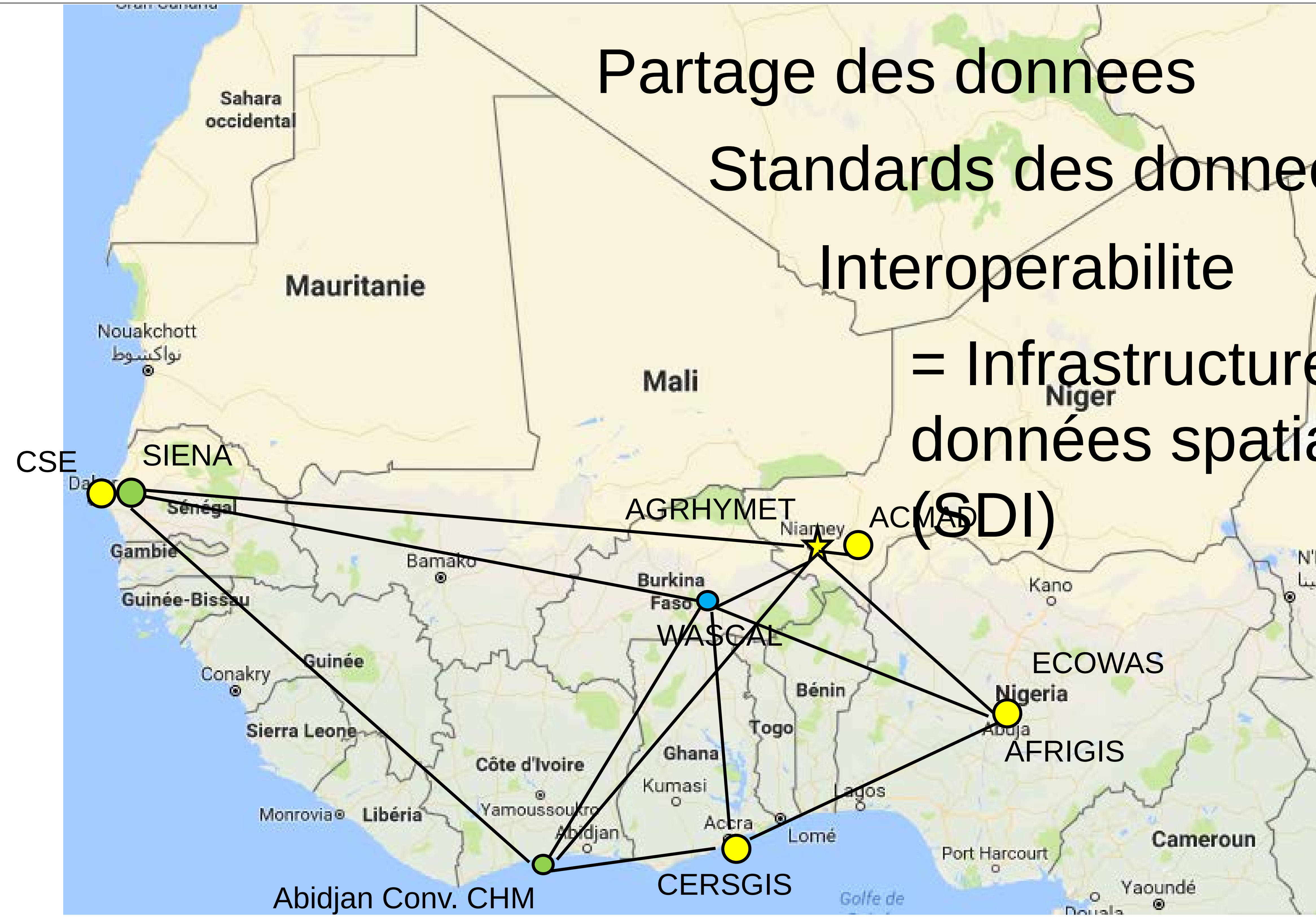
Vision

Partage des donnees

Standards des donnees

Interoperabilite

= Infrastructure des données spatiales (SDI)



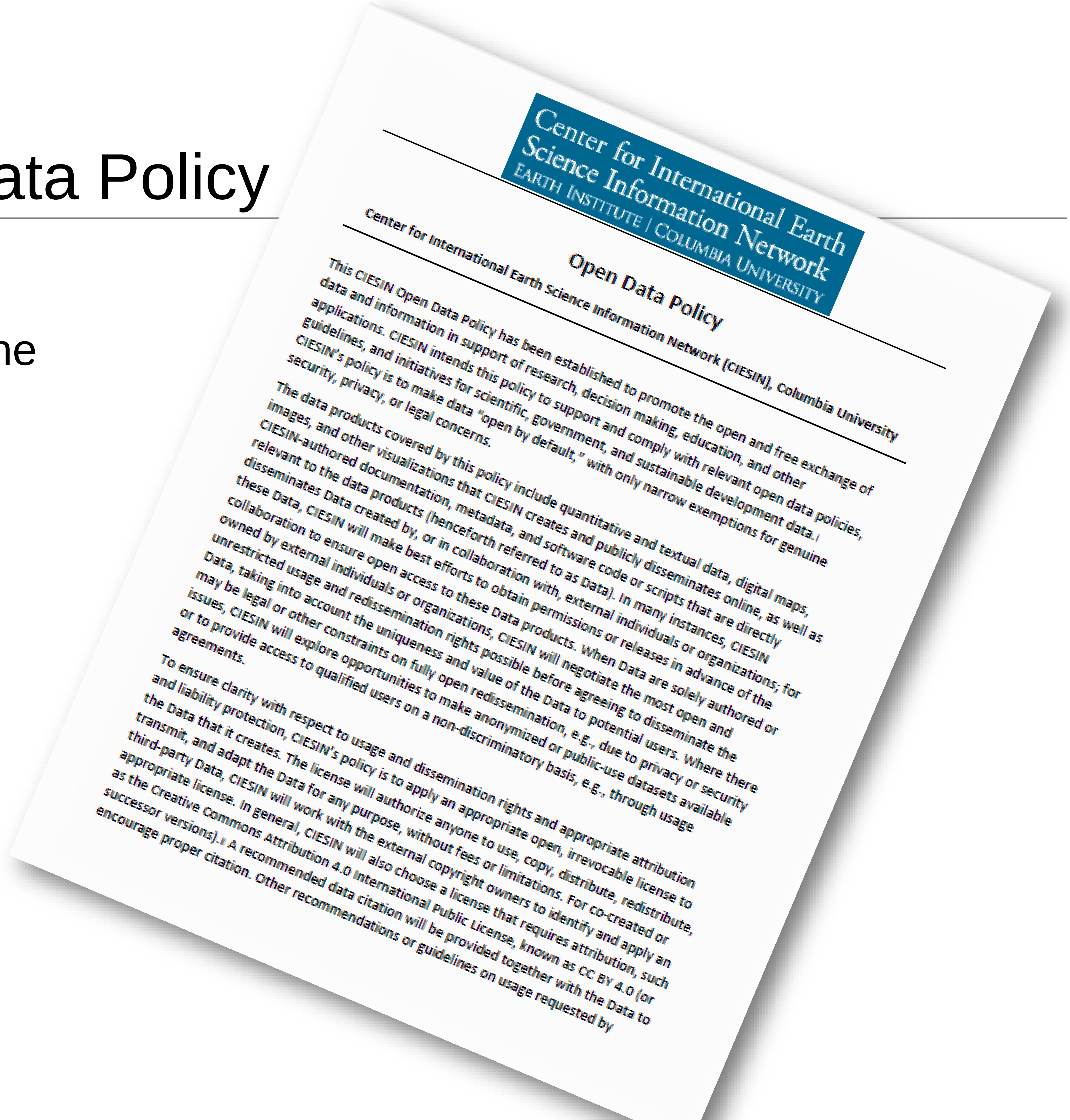
Data Policy

- Main issues if the (input) data come from somewhere else:
  - What is the provenance of the data?
  - Who owns the data (holds copyright)?
  - What restrictions have they placed on the data?
  - *It is best to obtain a signed permissions form from data providers*
- Main issues in setting data use constraints
  - Does your organization have an overarching data policy?
  - Are any of the data sensitive in nature (e.g., is there personally identifiable information, or information that should be restricted)?
  - Do you want to contribute to advance science *and* development through open data?

There is a major move in the scientific community towards open data!

CIESIN Data Policy

Available online



Center for International Earth Science Information Network
EARTH INSTITUTE | COLUMBIA UNIVERSITY

Center for International Earth Science Information Network (CIESIN), Columbia University

Open Data Policy

This CIESIN Open Data Policy has been established to promote the open and free exchange of data and information in support of research, decision making, education, and other applications. CIESIN intends this policy to support and comply with relevant open data policies, guidelines, and initiatives for scientific, government, and sustainable development data. CIESIN's policy is to make data "open by default," with only narrow exemptions for genuine security, privacy, or legal concerns.

The data products covered by this policy include quantitative and textual data, digital maps, images, and other visualizations that CIESIN creates and publicly disseminates online, as well as CIESIN-authored documentation, metadata, and software code or scripts that are directly relevant to the data products (henceforth referred to as Data). In many instances, CIESIN disseminates Data created by, or in collaboration with, external individuals or organizations; for these Data, CIESIN will make best efforts to obtain permissions or releases in advance of the collaboration to ensure open access to these Data products. When Data are solely authored or owned by external individuals or organizations, CIESIN will negotiate the most open and unrestricted usage and redissemination rights possible before agreeing to disseminate the Data, taking into account the uniqueness and value of the Data to potential users. Where there may be legal or other constraints on fully open redissemination, e.g., due to privacy or security issues, CIESIN will explore opportunities to make anonymized or public-use datasets available or to provide access to qualified users on a non-discriminatory basis, e.g., through usage agreements.

To ensure clarity with respect to usage and dissemination rights and appropriate attribution and liability protection, CIESIN's policy is to apply an appropriate open, irrevocable license to the Data that it creates. The license will authorize anyone to use, copy, distribute, redistribute, transmit, and adapt the Data for any purpose, without fees or limitations. For co-created or third-party Data, CIESIN will work with the external copyright owners to identify and apply an appropriate license. In general, CIESIN will also choose a license that requires attribution, such as the Creative Commons Attribution 4.0 International Public License, known as CC BY 4.0 (or successor versions). A recommended data citation will be provided together with the Data to encourage proper citation. Other recommendations or guidelines on usage requested by

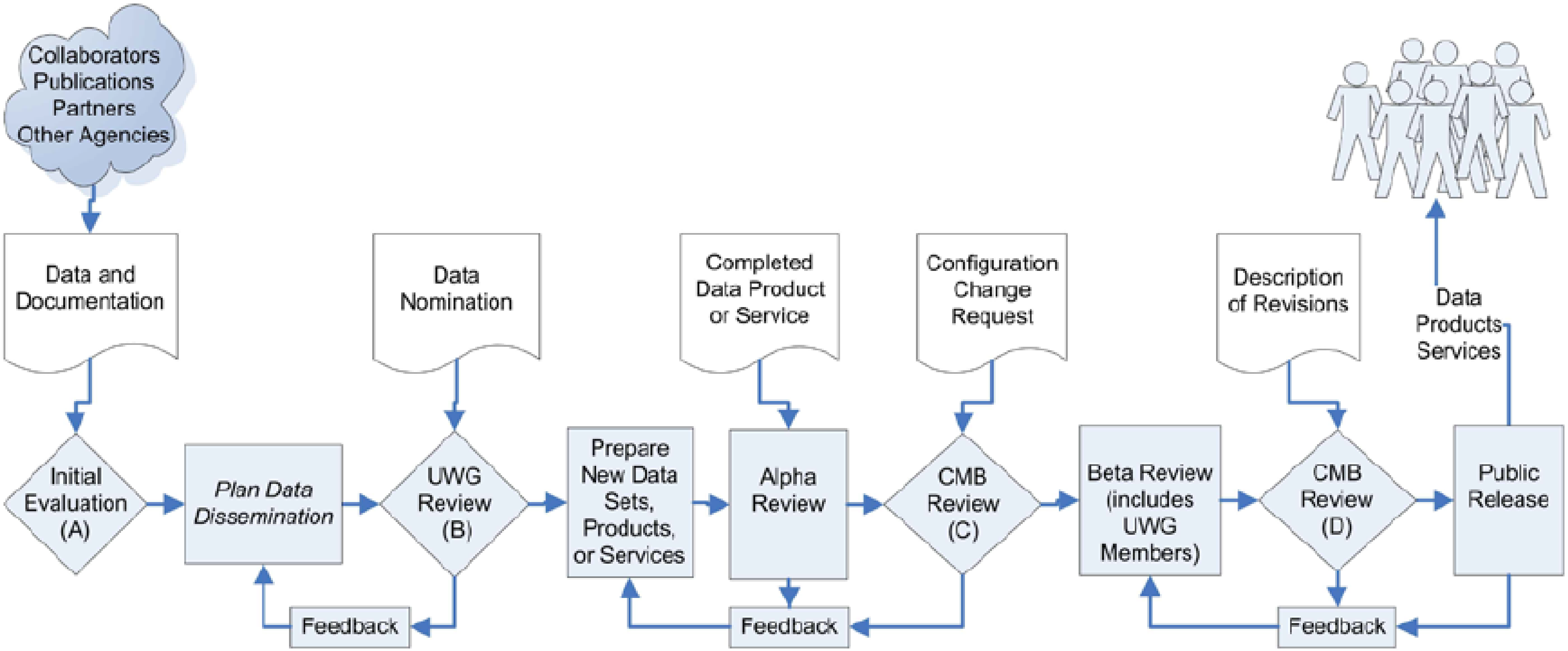
Choosing a License

- Creative Commons
  - https://creativecommons.org/share-your-work/licensing-types-examples/
  - Attribution (by)
  - ShareAlike (sa)
  - NonCommercial (nc)
  - NoDerivatives (nd)
  - No restrictions

"The ODC licenses apply only to sui generis database rights and any copyright in the **database structure**, they do not apply to the **individual contents of the database**. The latest version of the CC licenses on the other hand apply to sui generis **database rights and all copyright and neighboring rights in the database structure as well as the contents**."

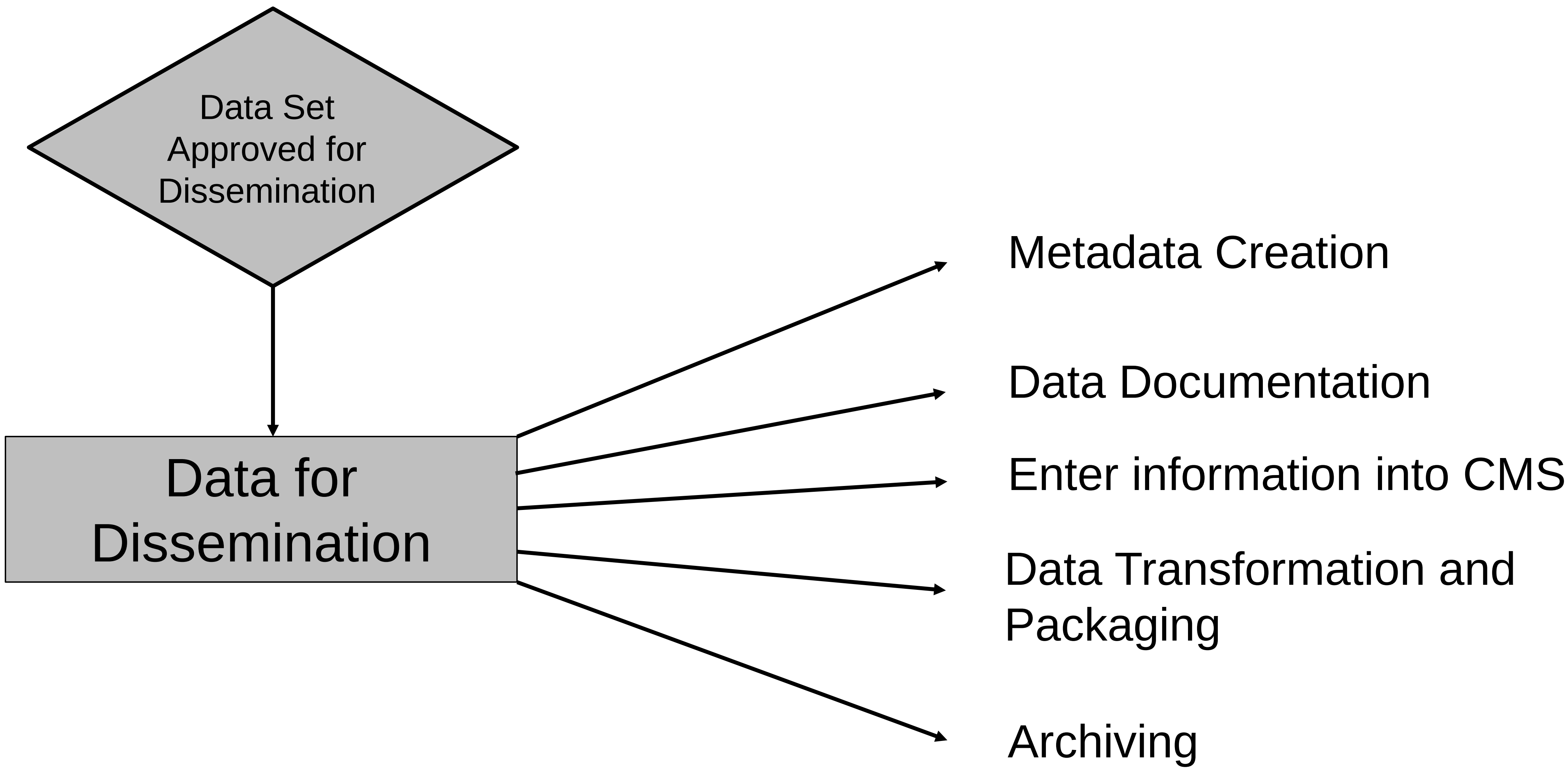
- Open Data Commons Licenses
  - https://opendatacommons.org/licenses/
  - Public Domain Dedication and License (PDDL) — “Public Domain for data/databases”
  - Attribution License (ODC-By) — “Attribution for data/databases”
  - Open Database License (ODC-ODbL) — “Attribution Share-Alike for data/databases”

Processus – de la curation jusqu'a la dissémination

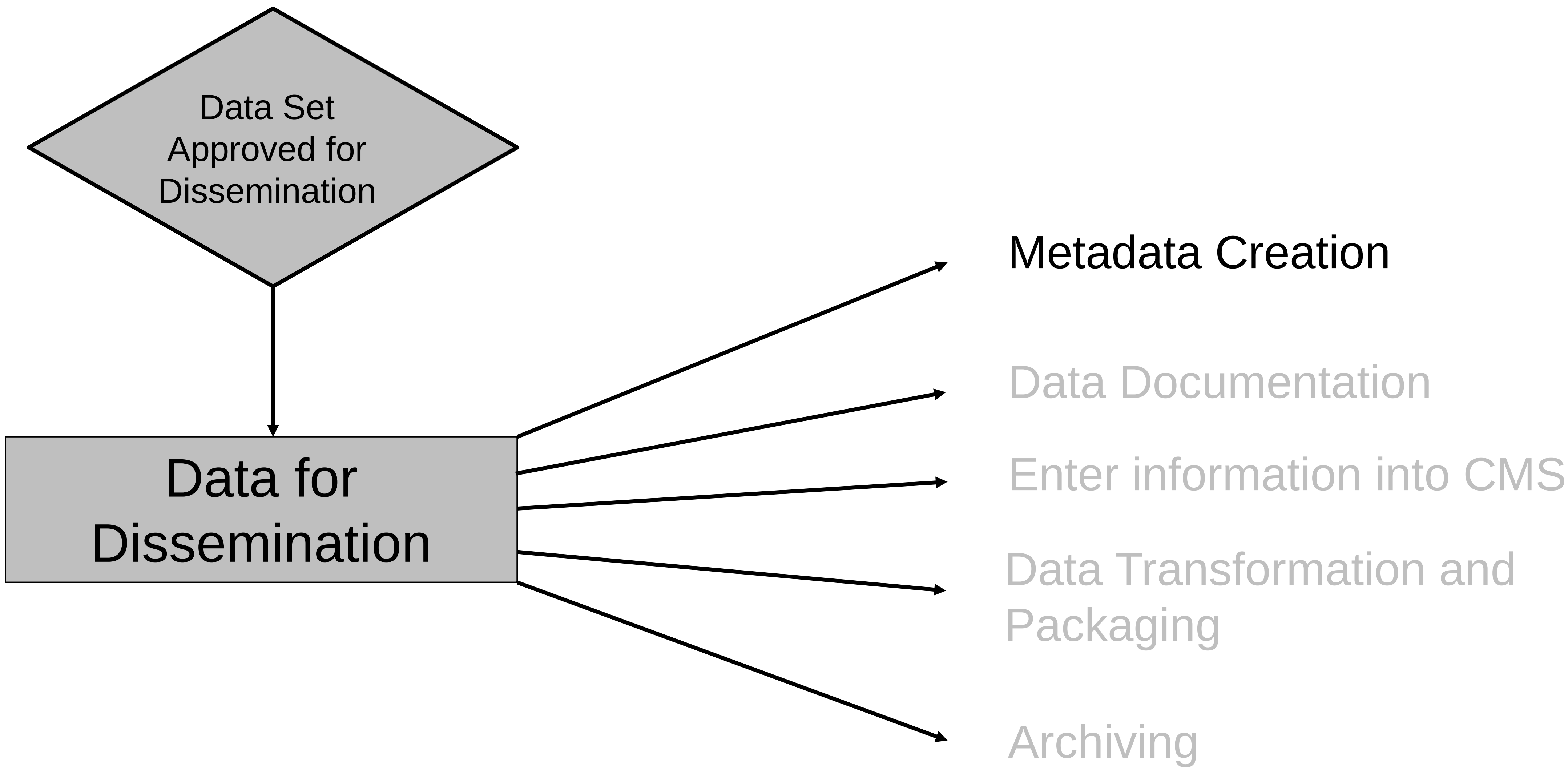


UWG = User Working Group
CMB = Configuration Management Board

Etapes dans la gestion et la dissémination des données

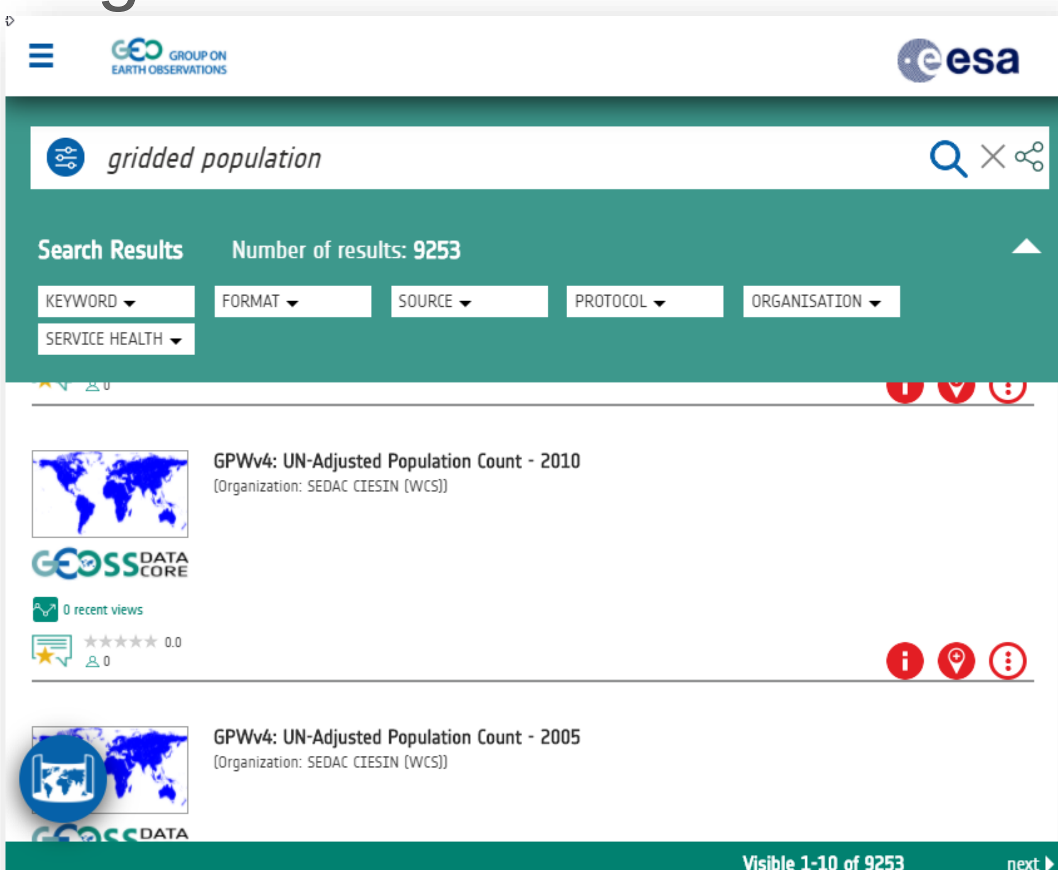


Etapes dans la gestion et la dissémination des données



Pourquoi developper metadata?

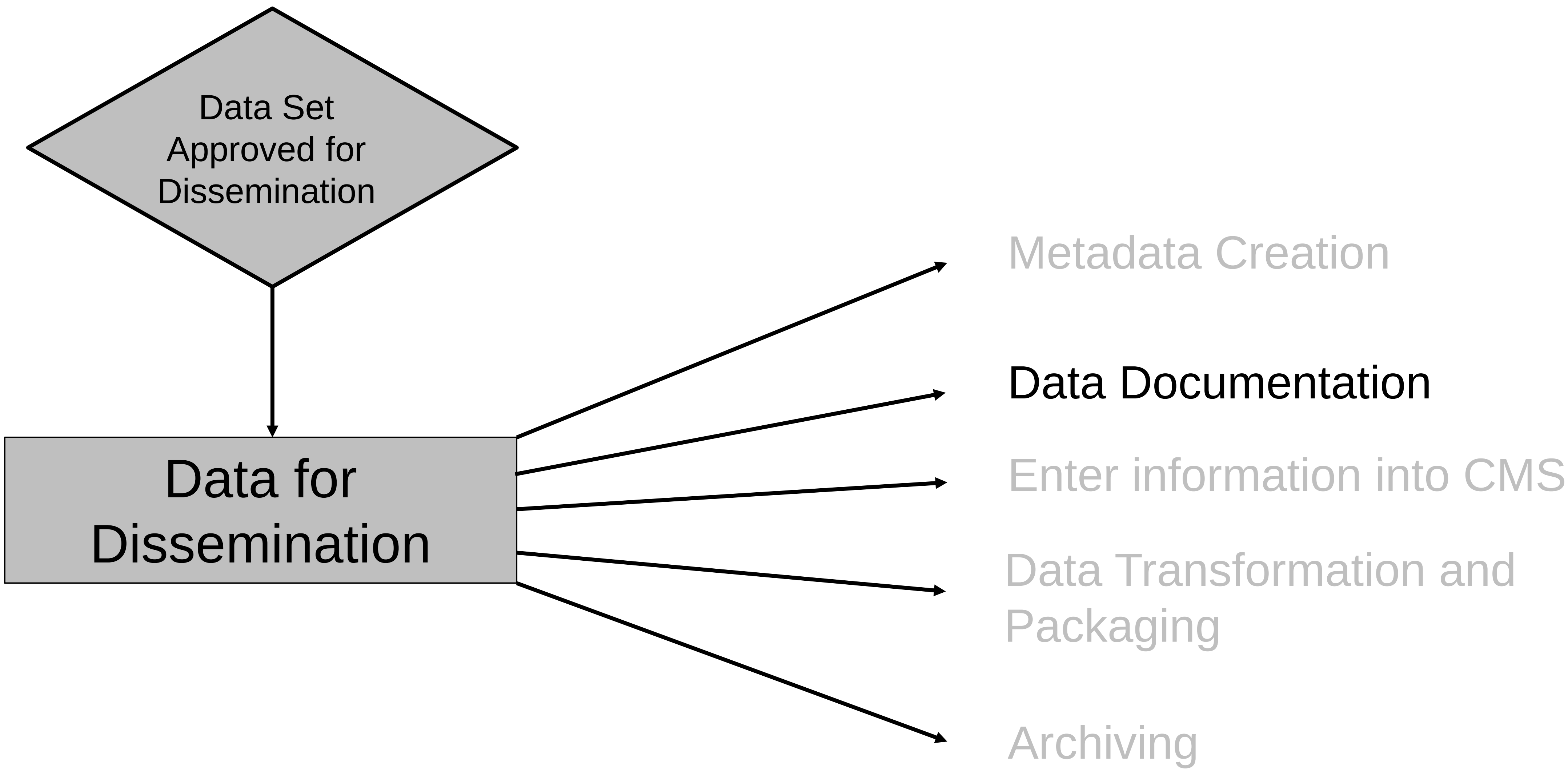
- Soutien la découverte des données
- Donne le contexte nécessaire pour utiliser et re-utiliser les données
- Quand ce sont standardisé, les metadata peuvent soutenir la récolte par les catalogues centraux, qui facilite une découverte élargie



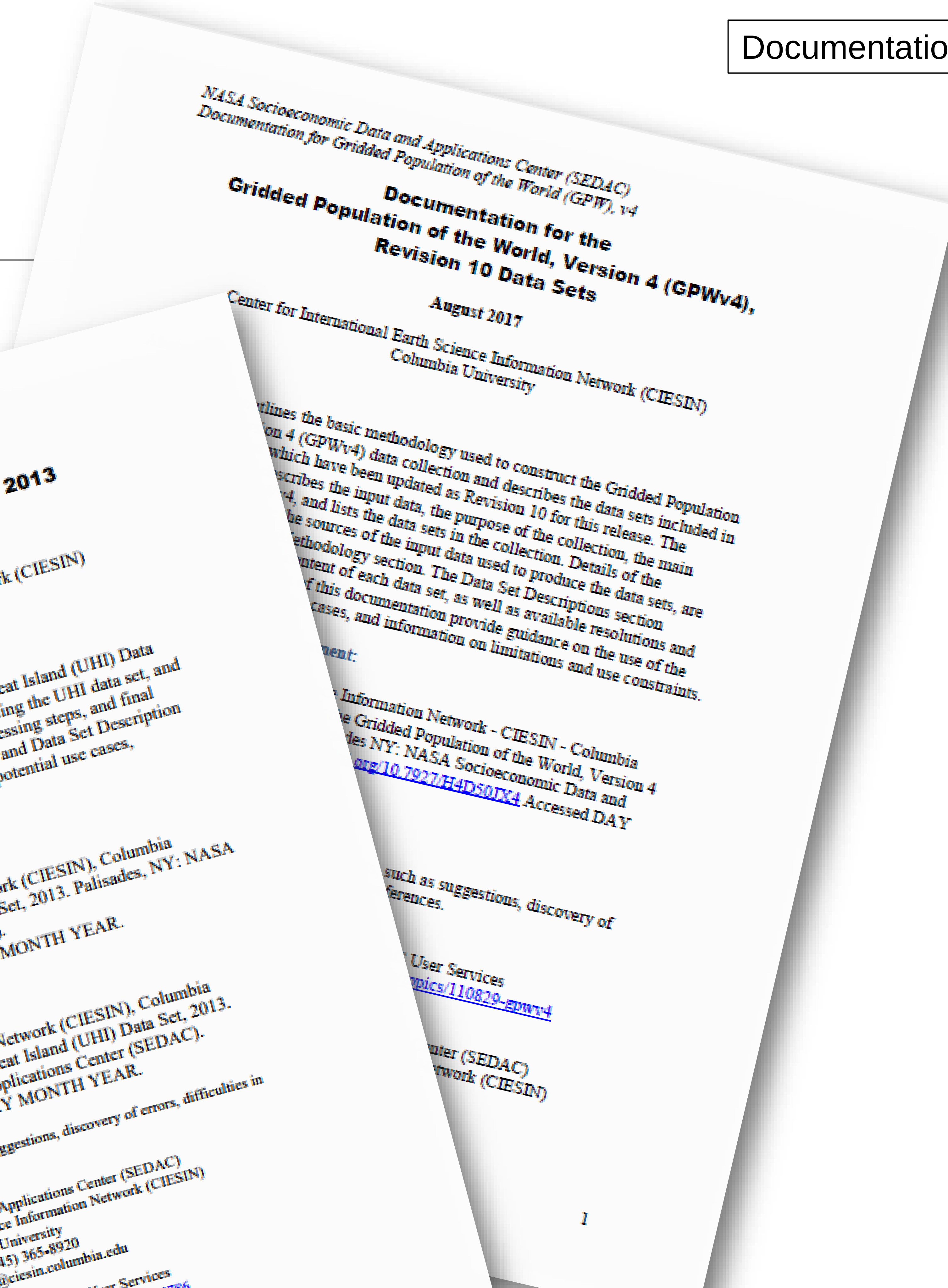
What is Metadata?

- More than just “data about data”
- **Data “reporting”**
  - **WHO** created the data?
  - **WHAT** is the content of the data?
  - **WHEN** was it created?
  - **WHERE** is it geographically?
  - **HOW** was the data developed?
  - **WHY** was the data developed?

Etapes dans la gestion et la dissémination des données



Documentation



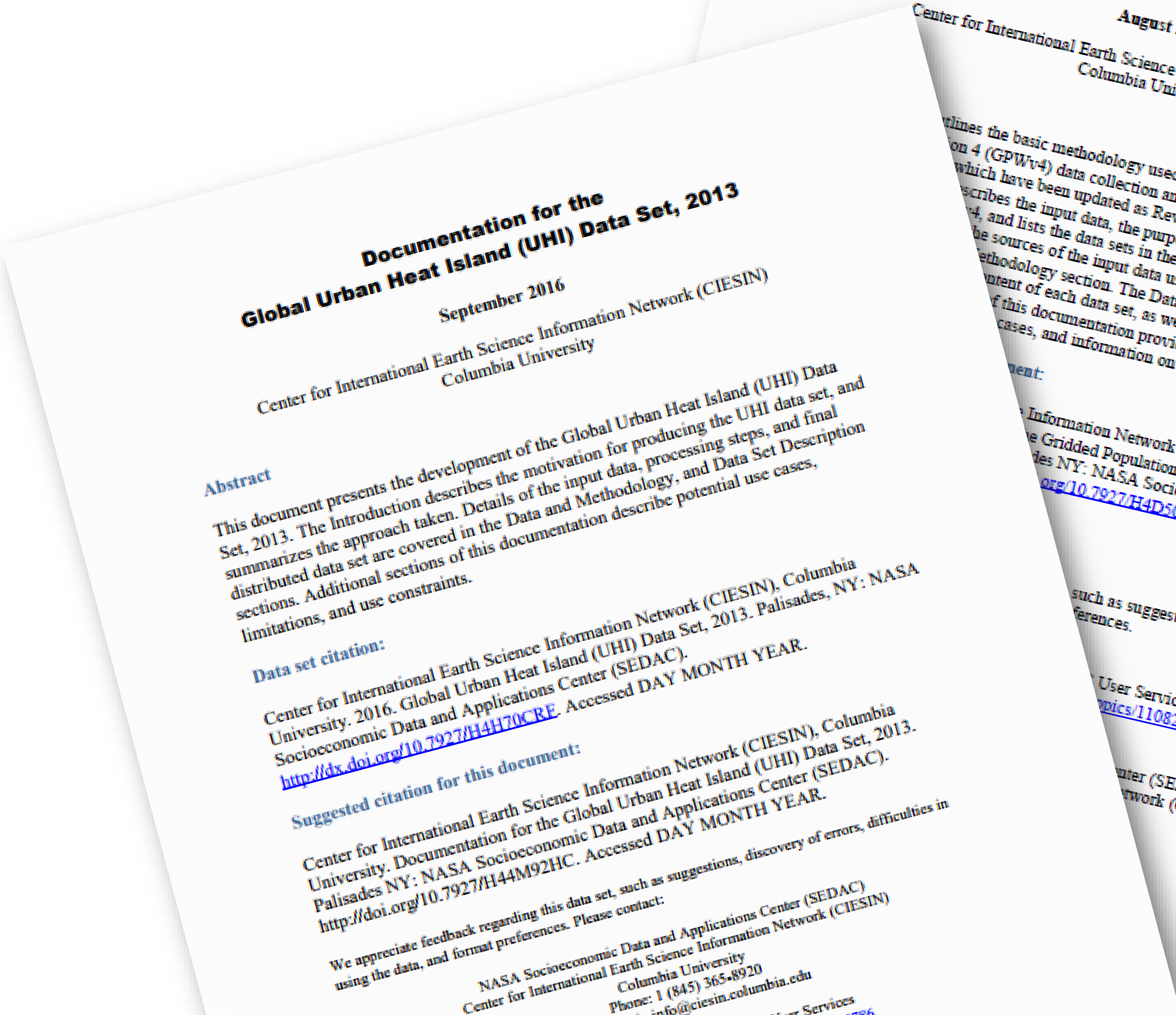
NASA Socioeconomic Data and Applications Center (SEDAC)
Documentation for Gridded Population of the World (GPW), v4

Documentation for the Gridded Population of the World, Version 4 (GPWv4), Revision 10 Data Sets

August 2017

Center for International Earth Science Information Network (CIESIN)
Columbia University

1



Documentation for the Global Urban Heat Island (UHI) Data Set, 2013

September 2016

Center for International Earth Science Information Network (CIESIN)
Columbia University

Abstract

This document presents the development of the Global Urban Heat Island (UHI) Data Set, 2013. The Introduction describes the motivation for producing the UHI data set, and summarizes the approach taken. Details of the input data, processing steps, and final distributed data set are covered in the Data and Methodology, and Data Set Description sections. Additional sections of this documentation describe potential use cases, limitations, and use constraints.

Data set citation:

Center for International Earth Science Information Network (CIESIN), Columbia University. 2016. Global Urban Heat Island (UHI) Data Set, 2013. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4H70CRF. Accessed DAY MONTH YEAR.

Suggested citation for this document:

Center for International Earth Science Information Network (CIESIN), Columbia University. Documentation for the Global Urban Heat Island (UHI) Data Set, 2013. Palisades NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://doi.org/10.7927/H44M92HC. Accessed DAY MONTH YEAR.

We appreciate feedback regarding this data set, such as suggestions, discovery of errors, difficulties in using the data, and format preferences. Please contact:

NASA Socioeconomic Data and Applications Center (SEDAC)
Center for International Earth Science Information Network (CIESIN)
Columbia University
Phone: 1 (845) 365-8920

Pour quoi la documentation?

- Documentation donne les détailles sur les méthodes et données utilises pour créer un base des données, les problèmes / échéances, et les exemples d'utilisation
- Documentation inadéquates peuvent représenter un obstacle a l'utilisation

ID	Date	Var1	Var2
1	9/1/16	33.7	35
2	9/2/16	22.424527	NA
3	9/3/16	22	-
4	9/4/16	55.66	-9999
5	9/5/16	1244.44	59
6	9/6/16	5.00E-08	66
7	9/7/16	44.5	42
8	9/8/16	756.32221	55

Quelles sont les problèmes avec ce base des données?

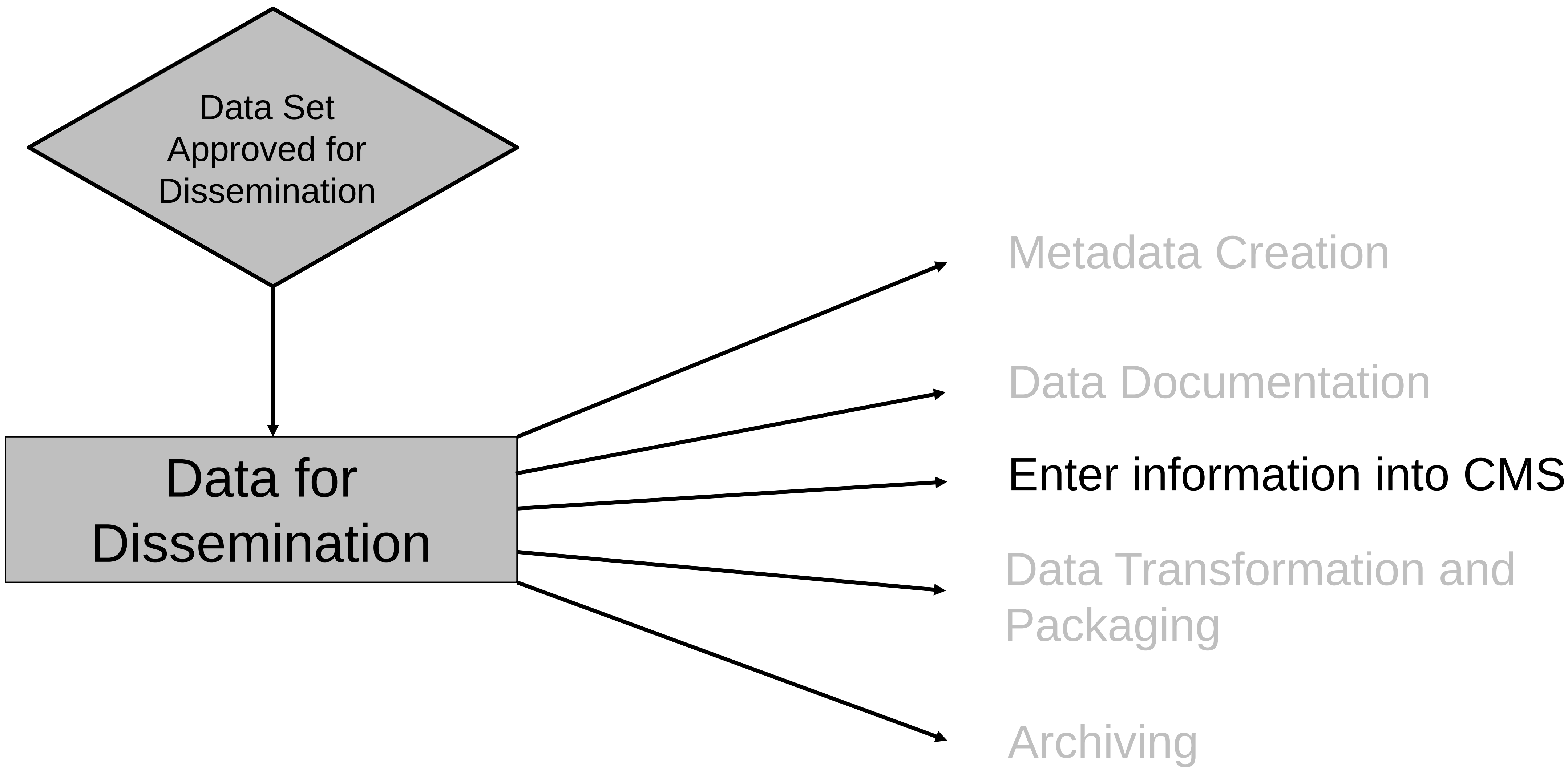
SEDAC Documentation: premiere page

- Documentation for <Dataset Title>
  <Documentation Publication Date>
  <Authors>
- Abstract
- Data set citation
- Suggested citation for documentation
- Contact to provide feedback on documentation

SEDAC Documentation: contenu

Enter into CMS

Etapes dans la gestion et la dissémination des données



Enter into CMS

Overview

Collection Overview

Methods

Data Sets (8)

Population Density, v4 (2000, 2005, 2010, 2015, 2020)

Show All...

Map Gallery (27)

Map Services (26)

Citations

FAQs

What's New in GPWv4?

Documentation

What is UN-adjusted population data?

Multimedia

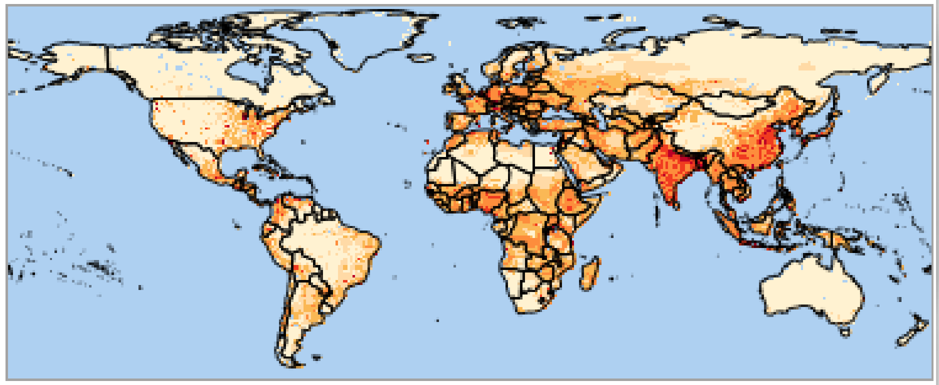
Acknowledgments

SEDAC Hazards

Population Density, v4 (2000, 2005, 2010, 2015, 2020)

Set Overview | Data Download | Maps | Map Services | Documentation | Metadata

GPWv4: Population Density - 2015



4 of 5

Purpose:

To provide estimates of population density for the years 2000, 2005, 2010, 2015, and 2020, based on counts consistent with national censuses and population registers, as raster data to facilitate data integration.

Abstract:

Gridded Population of the World, Version 4 (GPWv4) Population Density consists of estimates of human population density based on counts consistent with national censuses and population registers, for the years 2000, 2005, 2010, 2015, and 2020. A proportional allocation gridding algorithm, utilizing approximately 12.5 million national and sub-national administrative units, is used to assign population values to 30 arc-second (~1 km) grid cells. The population density grids are created by dividing the population count grids by the land area grids. The pixel values represent persons per square kilometer.

Recommended Citation(s)*:

Center for International Earth Science Information Network - CIESIN - Columbia University. 2016. Gridded Population of the World, Version 4 (GPWv4): Population Density. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4NP22DQ. Accessed DAY MONTH YEAR

ENW (EndNote & RefWorks)†

RIS (Others)

feedback and support

Enter into CMS

Data Download

Collection Overview

Methods

Data Sets (8)

Population Density, v4 (2000, 2005, 2010, 2015, 2020)

Show All...

Map Gallery (27)

Map Services (26)

Citations

FAQs

What's New in GPWv4?

Documentation

What is UN-adjusted population data?

Multimedia

Acknowledgments

Population Density, v4 (2000, 2005, 2010, 2015, 2020)

Set Overview | Data Download | Maps | Map Services | Documentation | Metadata

Downloads

Data:

View Recommended Citation(s)

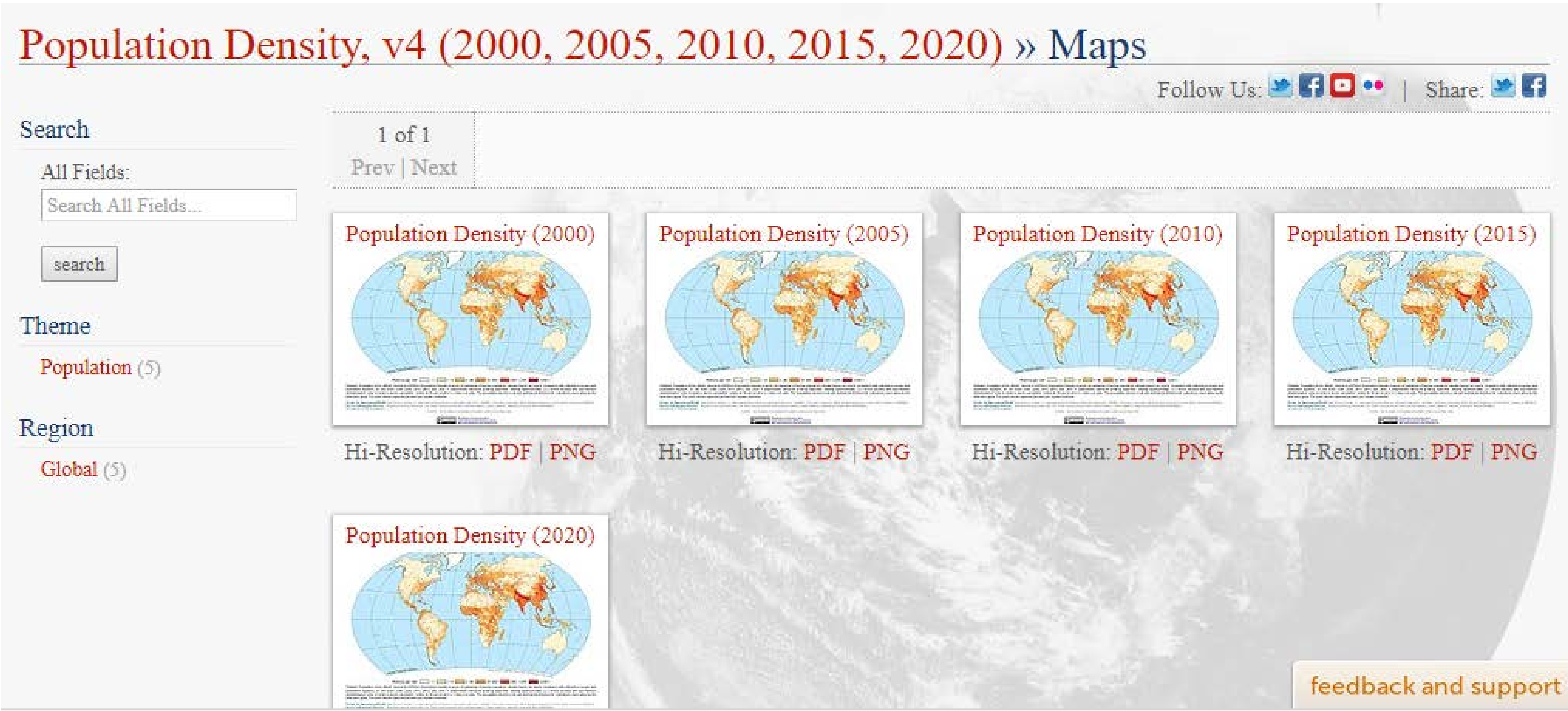
Note: For regional to global analyses, users may wish to download the UN-adjusted versions of this data set. Further explanations as to the differences between the non-adjusted and UN-adjusted versions of GPWv4 are found on the What is UN-Adjusted data? web page.

Gridded Population of the World, Version 4 (GPWv4): Population Density are available as global grids in GeoTiff format. Each downloadable is a compressed zip file, which contains: 1) the global GeoTiff for the year of estimate, 2) PDF documentation, 3) a Microsoft Excel file (.xlsx) with country-level information and sources, and 4) a text file (.txt) with a log of changes to the dataset by version.

Year of Estimate | 2000 2005 2010 2015 2020 (each download is ~180 MB)

Enter into CMS

Map Gallery



Enter into CMS

Map Services

- Collection Overview
- Methods
- Data Sets (8)
  - *Population Density, v4 (2000, 2005, 2010, 2015, 2020)*
  - ⊞ Show All...
- Map Gallery (27)
- Map Services (26)
- Citations
- FAQs
- What's New in GPWv4?
- Documentation
- What is UN-adjusted population data?
- Multimedia
- Acknowledgments
- SEDAC Hazards Mapper
- Population Estimation

Population Density, v4 (2000, 2005, 2010, 2015, 2020)

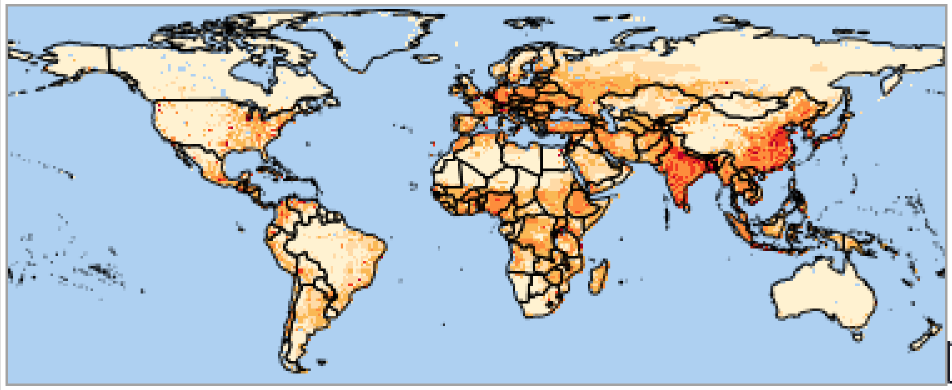
Set Overview | Data Download | Maps | Map Services | Documentation | Metadata

Clicking on the map widget to the right will launch an interactive map tool which will allow you to zoom, pan and switch layers.

Here's an example on how to create a WMS layer instance using the Open Source JavaScript library OpenLayers.

```
var wms = new OpenLayers.Layer.WMS(
"Population Density",
"http://sedac.ciesin.columbia.edu/geoserver/wms",
{layers: 'gpw-v3:gpw-v3-population-density_2000'}
);
```

GPWv4: Population Density - 2005



2 of 5

The possible values for `layers` can be found in the list below.

GPWv4: Population Density - 2000	gpw-v4:gpw-v4-population-density_2000
GPWv4: Population Density - 2005	gpw-v4:gpw-v4-population-density_2005
GPWv4: Population Density - 2010	gpw-v4:gpw-v4-population-density_2010
GPWv4: Population Density - 2015	gpw-v4:gpw-v4-population-density_2015
GPWv4: Population Density - 2020	gpw-v4:gpw-v4-population-density_2020

Enter into CMS

Documentation

- Collection Overview
- Methods
- Data Sets (8)
  - *Population Density, v4 (2000, 2005, 2010, 2015, 2020)*
  - Show All...
- Map Gallery (27)
- Map Services (26)
- Citations
- FAQs
- What's New in GPWv4?
- Documentation
- What is UN-adjusted population data?
- Multimedia
- Acknowledgments
- SEDAC Hazards Mapper

Population Density, v4 (2000, 2005, 2010, 2015, 2020)

Set Overview | Data Download | Maps | Map Services | Documentation | Metadata

Data Collection Documentation:

- GPWv4 documentation (PDF)
- Country-level Information and Sources (Microsoft Excel .xlsx file)
- Log of changes to the dataset by version

Additional Documentation:

- Detailed descriptions of the methods and improvements made in the GPWv4 data collection are described in the following paper by Doxsey-Whitfield et al. (2015): Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4 (GPWv4)
- NASA EarthData Webinar: Discover NASA's Updated Gridded Population of the World Data, January 2015 (1 hour long)

Enter into CMS

Metadata

Collection Overview
Methods
Data Sets (8)
Population Density, v4 (2000, 2005, 2010, 2015, 2020)
⊞ Show All...
Map Gallery (27)
Map Services (26)
Citations
FAQs
What's New in GPWv4?
Documentation
What is UN-adjusted population data?
Multimedia
Acknowledgments
SEDAC Hazards Mapper
Population Estimation

Population Density, v4 (2000, 2005, 2010, 2015, 2020)

Set Overview | Data Download | Maps | Map Services | Documentation | Metadata

Browse Metadata

Identification | Data Quality | Spatial Data Organization | Spatial Reference | Entity and Attribute | Distribution | Metadata Reference | File Formats: XML, HTML, Text

Identification Information:

Citation:

Citation Information:

Originator: Center for International Earth Science Information Network - CIESIN - Columbia University
Publication Date: 2016
Publication Time:
Title:

Gridded Population of the World, Version 4 (GPWv4): Population Density

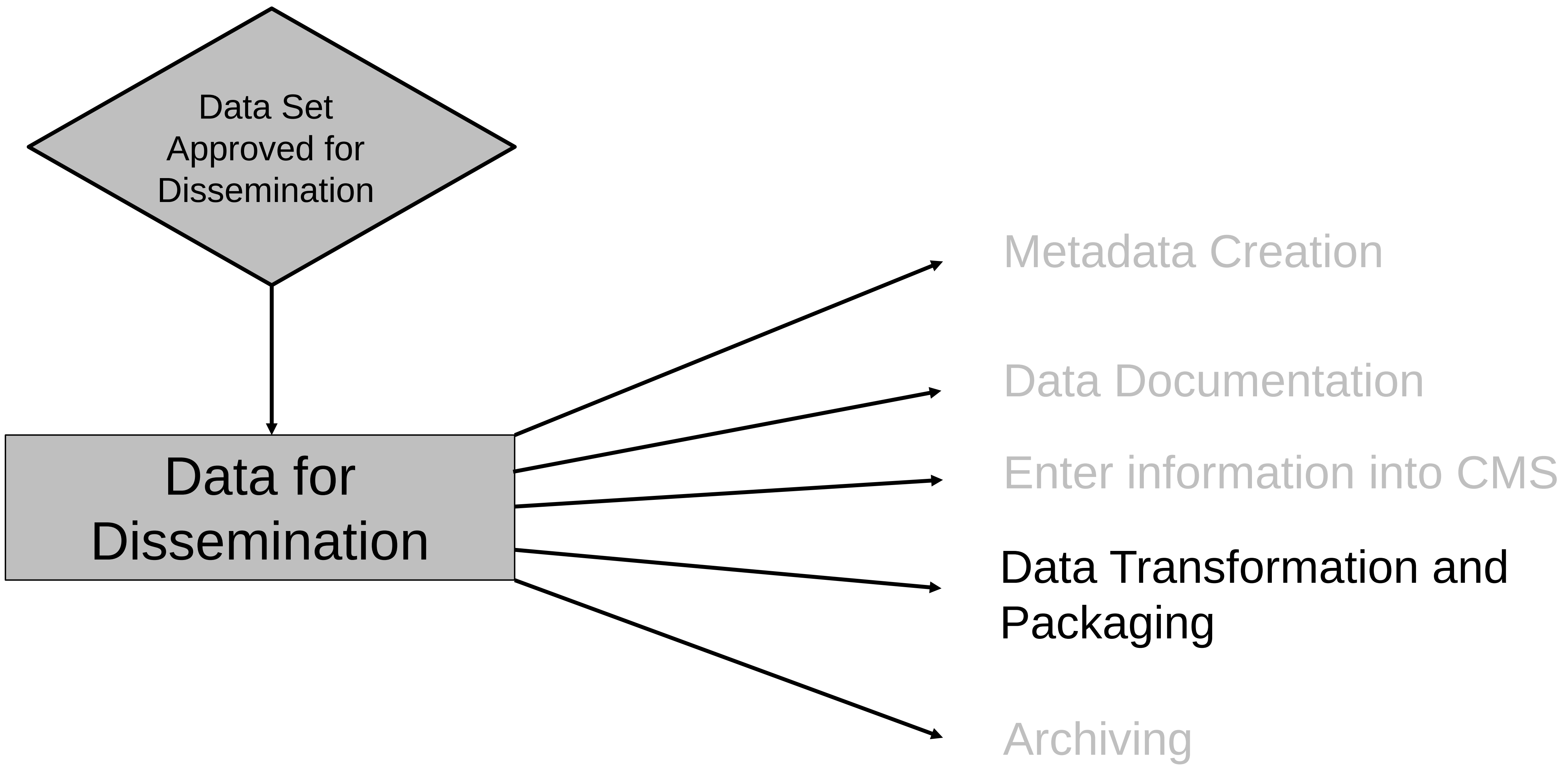
Edition: 4.00
Geospatial Data Presentation Form: raster, map
Series Information:

Series Name:
Issue Identification:

Publication Information:

Publication Place: Palisades, NY
Publisher: NASA Socioeconomic Data and Applications Center (SEDAC)

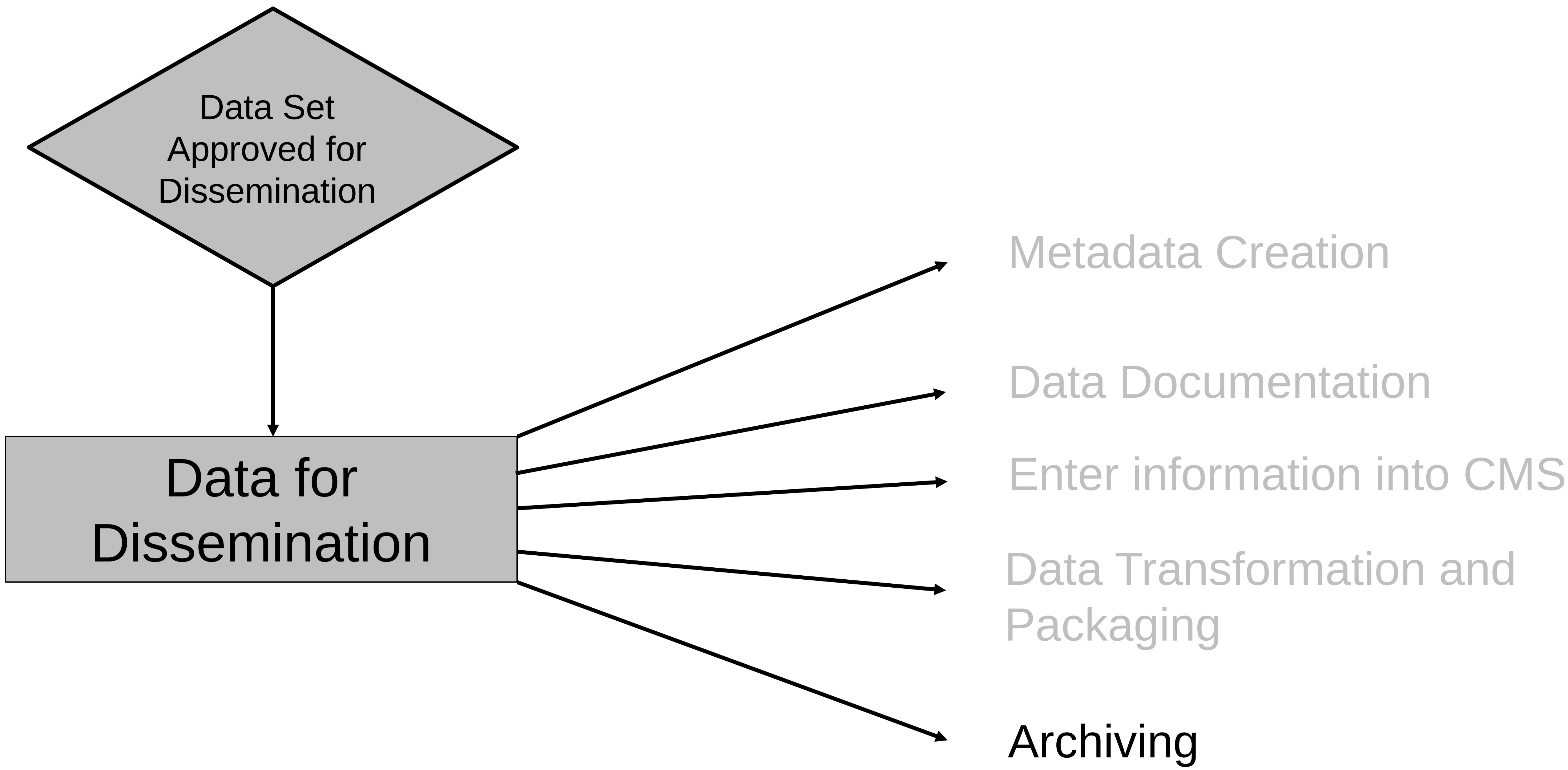
Etapes dans la gestion et la dissémination des données



Transformation et l’emballage

- Convert to different projections (generally Geographic)
- Convert to different resolutions
- Subset data (e.g., regional / continental subsets)
- Convert to different formats (e.g., Esri GRID or ASCII to GeoTIFF)
- Create thumbnail and map gallery maps
- Zip the data *with* documentation

Etapes dans la gestion et la dissémination des données



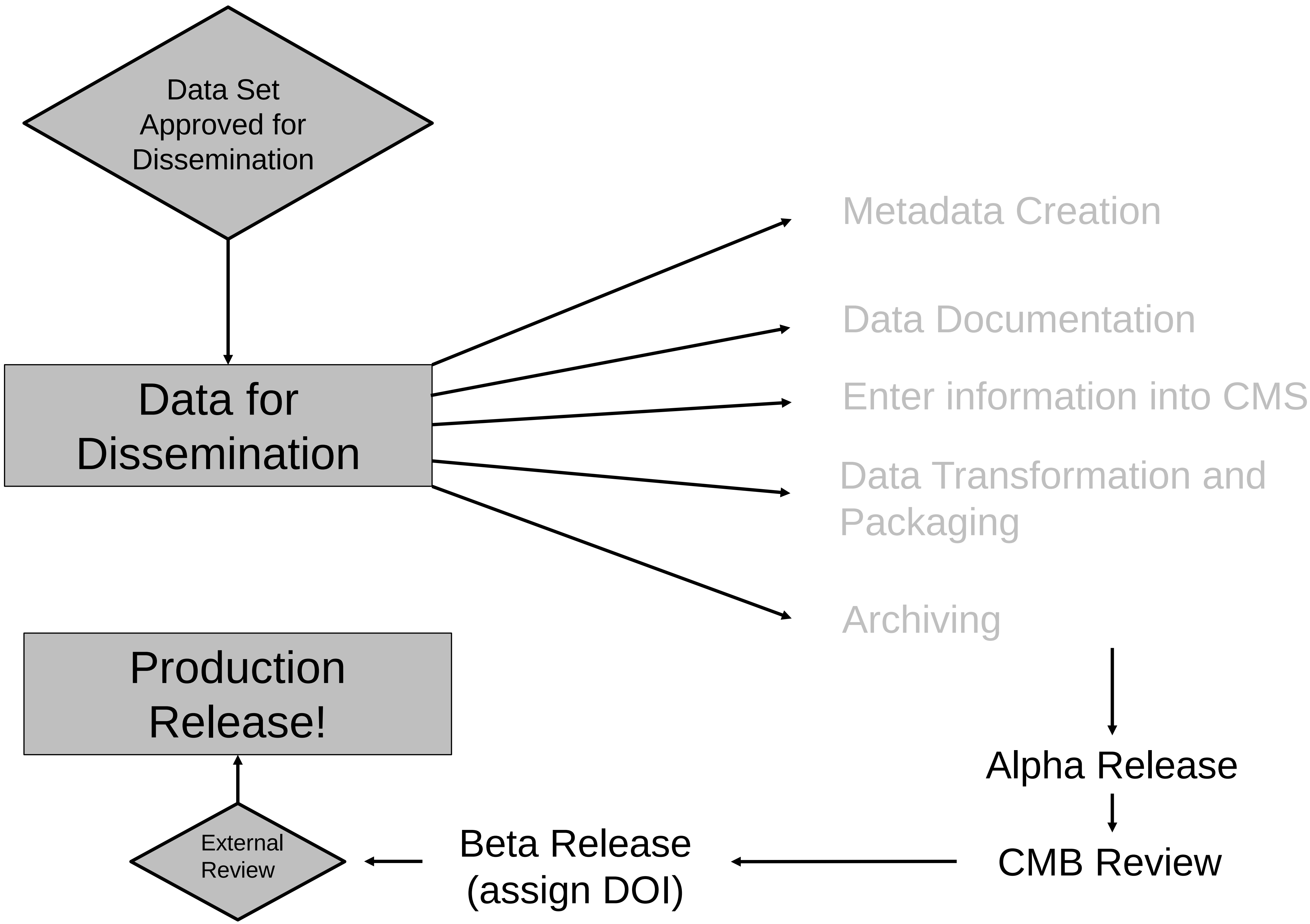
Archiving (1)

- Submit Archive Ingest Form
- Create a new archival metadata record in the Tracking Database (TDB)
- Archival material package may include:
  - Archives Ingest forms
  - Permissions forms
  - E-mails related to the archival process
  - Superseded or older versions of files
  - Compressed original files
  - Preserved Web pages/sites
  - Metadata records
  - Recommended citations
  - Digital scans of relevant non-digital documents
  - Any other related documents or files

Archiving (2)

- Use “BagIt” software from the Library of Congress (https://github.com/LibraryOfCongress/bagit-java) to package the data for archiving
  - This creates a zip file for the “content” and a zip file for the “nonpublic” parts of the data set
- Burn Optical Disc Media
  - Once the “bags” have been created, they should be copied to industry-standard, write-once, optical media
- Verify disks, run anti-virus, and label disks
- Store labeled disks in secure, climate controled room

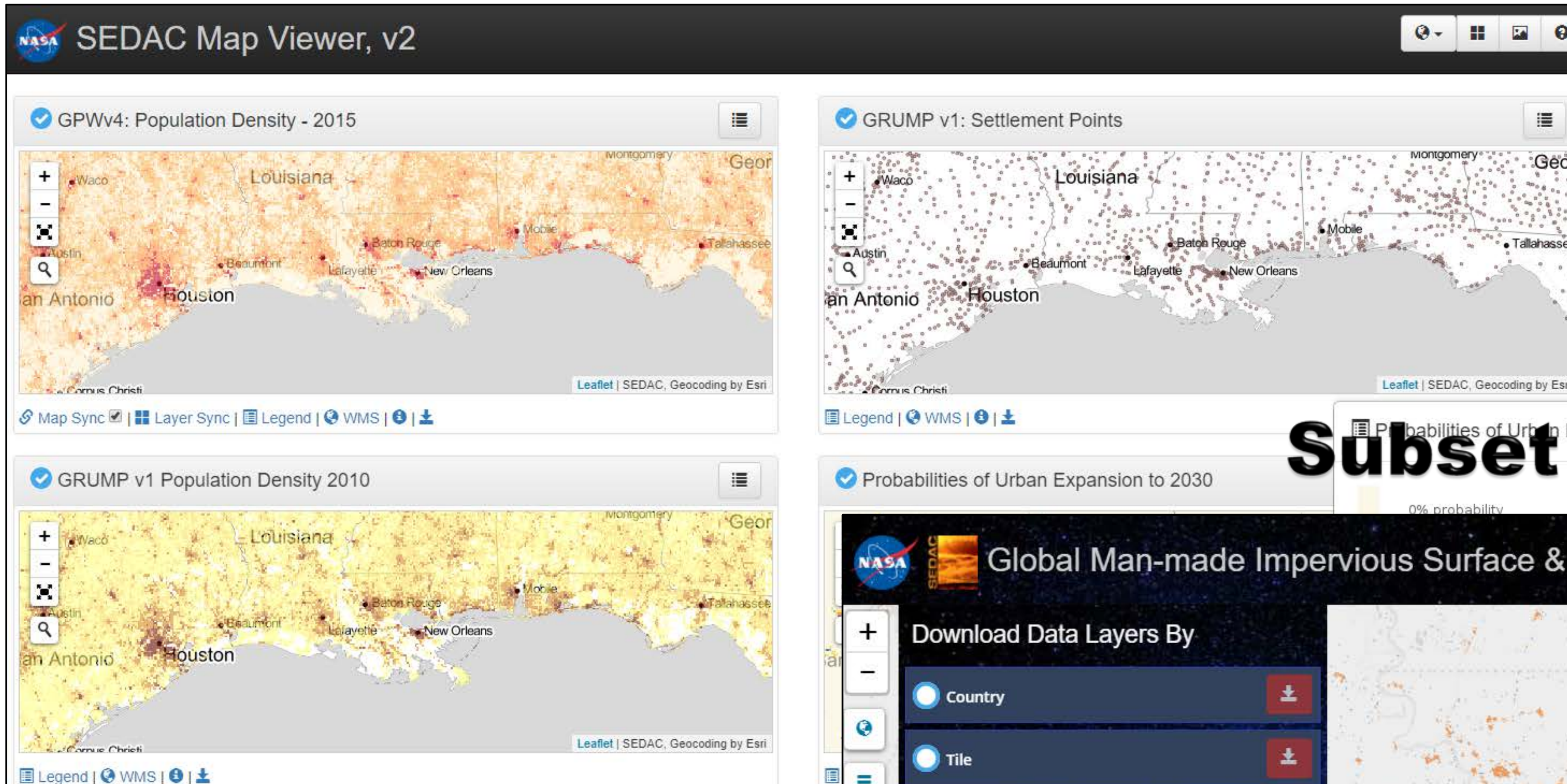
Etapes dans la gestion et la dissémination des données



Et beaucoup d’autres choses…

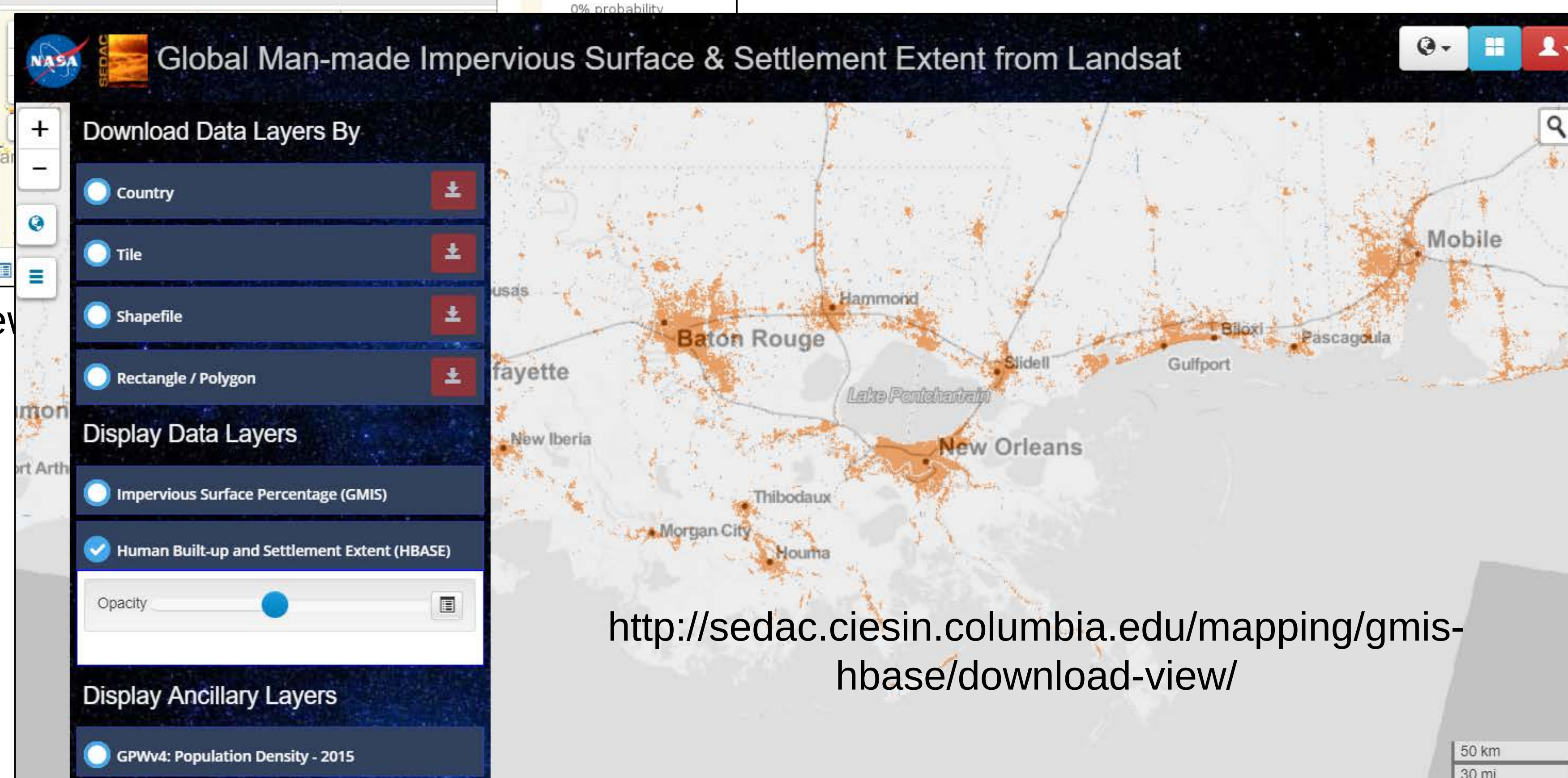
- Infrastructure ou « IT »
  - Servers
  - Sécurité
- Base des connaissances scientifique et technique
  - Présent dans la communauté scientifique
    - Relations avec des utilisateurs des données
    - Relations avec des détenteurs/trices de données
  - Connaissances en matière de l’informatique
  - Connaissances en matière des sciences de l’information / archivage
- Processus bien défini
  - DST
  - CMB
- Développement des services basées sur les données
  - Programmeurs

Intercomparison Tool



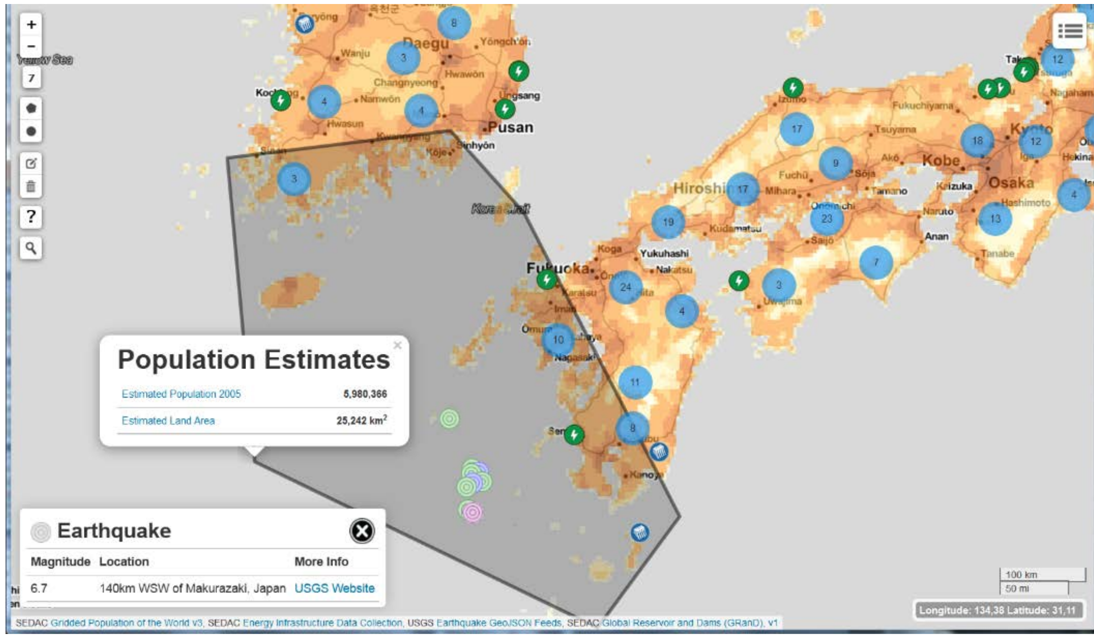
http://sedac.ciesin.columbia.edu/mapping/vie

Subset & Download Tool



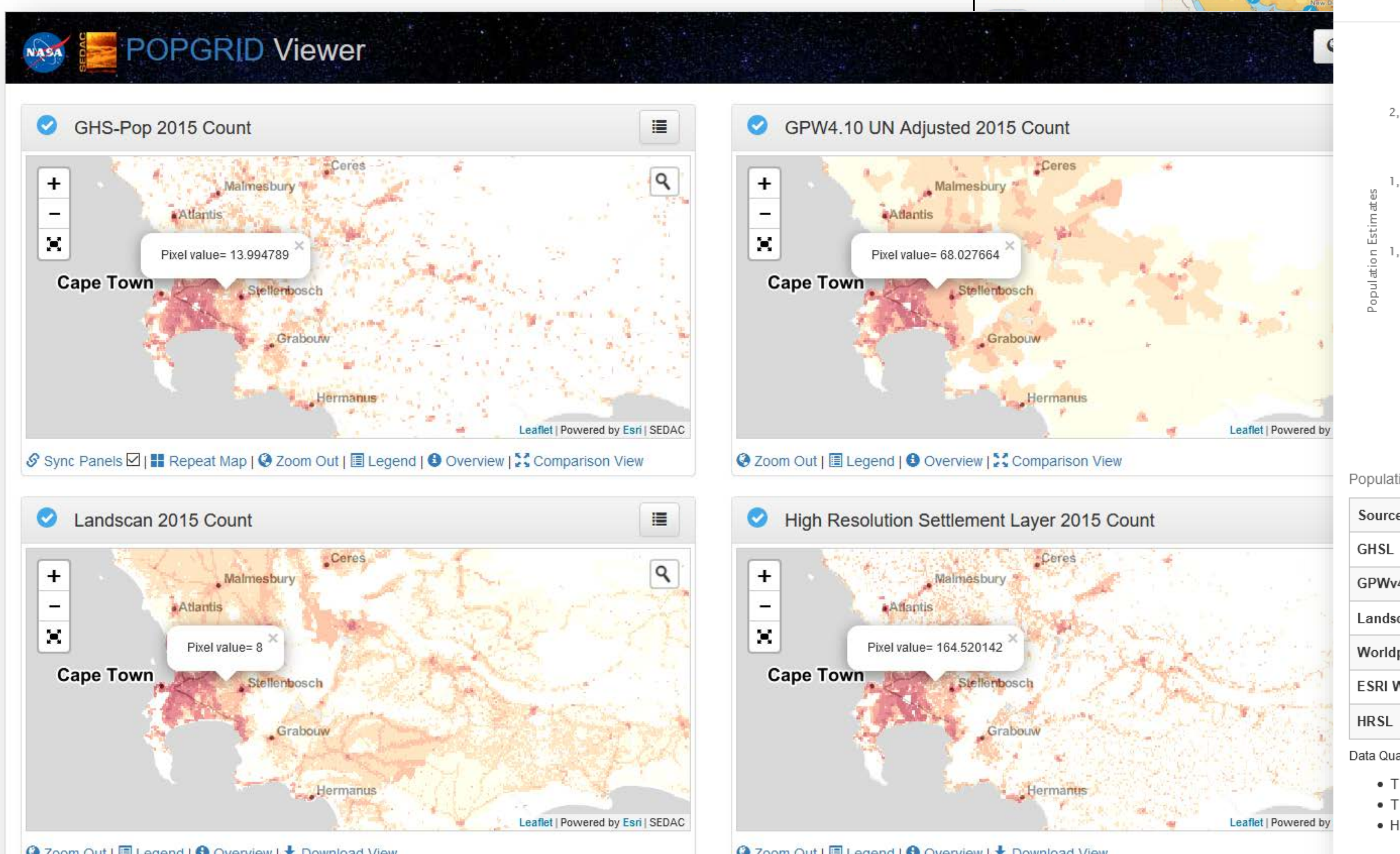
http://sedac.ciesin.columbia.edu/mapping/gmis-hbase/download-view/

Population Estimation Service



HazPop App

POPGRID Viewer



Merci de votre attention!

adesherbinin@ciesin.columbia.edu

www.ciesin.columbia.edu

NASA SEDAC:

http://sedac.ciesin.columbia.edu

ICSU WORLD DATA SYSTEM

DISCUSSION

If all the data are on the cloud do we still need repositories?

- Many data centers are moving their repositories to the cloud
- Cloud services can save money, especially for storage, back up, and security
- Guarantees up-to-date infrastructure and scalability
- It implies a different cost model
  - Instead of spending money on fixed infrastructure, costs are monthly
  - Costs are based partly on storage and partly on egress (getting data out)
- Cloud services do not eliminate the need for domain expertise
  - For curation
  - For management

Domain or Open Repositories?

Domain Specific

- Expertise for curation and management
- Commitment to long-term preservation
- Expert guidance by advisory groups
- Links to larger networks
- Higher likelihood of data discovery

Examples: WDS members

Open Repositories

- Lower costs of operation
- Lower level of effort on the part of researchers
- Satisfies the requirement for data availability by journals and some funders

Examples: Zenodo, Mendelay, Dryad, etc.

An archive is more than preserving bits and bytes… it needs to enable people to use the data set in the future

Feuille de route en Afrique de l’Ouest

- Est-ce que les entrepôts de données en Afriques sont nécessaires?
- Si oui, a quel niveau: continental (sous l’egide de l’AOSP ou OAU), régionale, ou nationale?
- Si oui, entrepôts pour des différents domaines scientifiques ou entrepôts générique?
- Quels sont les modes de financement?
  - Souscription des pays
  - Bailleurs de fonds
- Quels sont les capacités techniques? Comment les renforcer?

BACKUP SLIDES

CoreTrustSeal Self Assessment

- CoreTrustSeal is a new organization merging certification processes from the ICSU WDS and the Data Seal of Approval (DSA)
- It is a certification body that helps data centers conform to best practices
- The self assessment addresses key elements and best practices of data management
- The full requirements are at http://www.coretrustseal.org

An archive is more than preserving bits and bytes… it needs to enable people to use the data set in the future

Organisational infrastructure

- ~~R1. The repository has an explicit mission to provide access to and~~ preserve data in its domain.
- R2. The repository maintains all applicable licenses covering data access and use and monitors compliance.
- R3. The repository has a continuity plan to ensure ongoing access to and preservation of its holdings.
- R4. The repository ensures, to the extent possible, that data are created, curated, accessed, and used in compliance with disciplinary and ethical norms.
- R5. The repository has adequate funding and sufficient numbers of qualified staff managed through a clear system of governance to effectively carry out the mission.
- R6. The repository adopts mechanism(s) to secure ongoing expert guidance and feedback (either in-house, or external, including scientific guidance, if relevant).

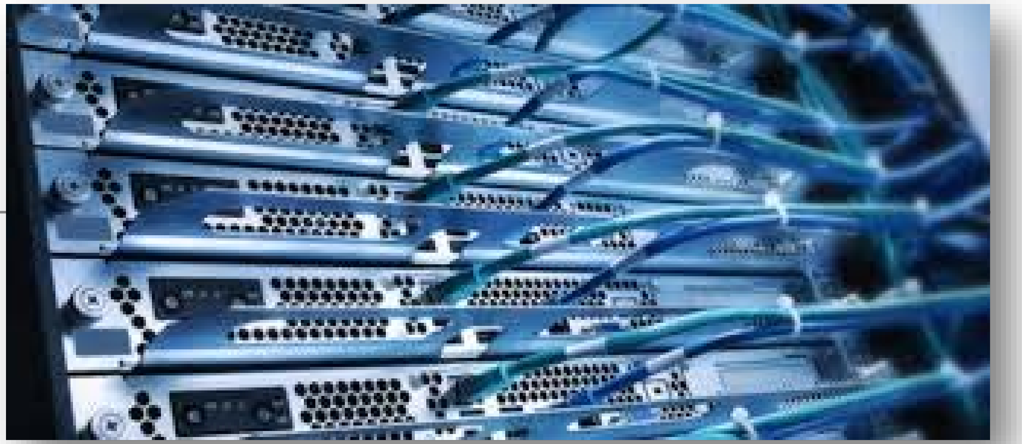
Digital object management

- R7. The repository guarantees the integrity and authenticity of the data.
- R8. The repository accepts data and metadata based on defined criteria to ensure relevance and understandability for data users.
- R9. The repository applies documented processes and procedures in managing archival storage of the data.
- R10. The repository assumes responsibility for long-term preservation and manages this function in a planned and documented way.

Digital object management

- R11. The repository has appropriate expertise to address technical data and metadata quality and ensures that sufficient information is available for end users to make quality-related evaluations.
- R12. Archiving takes place according to defined workflows from ingest to dissemination.
- R13. The repository enables users to discover the data and refer to them in a persistent way through proper citation.
- R14. The repository enables reuse of the data over time, ensuring that appropriate metadata are available to support the understanding and use of the data.

Technical infrastructure



- R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.
- R16. The technical infrastructure of the repository provides for protection of the facility and its data, products, services, and users.