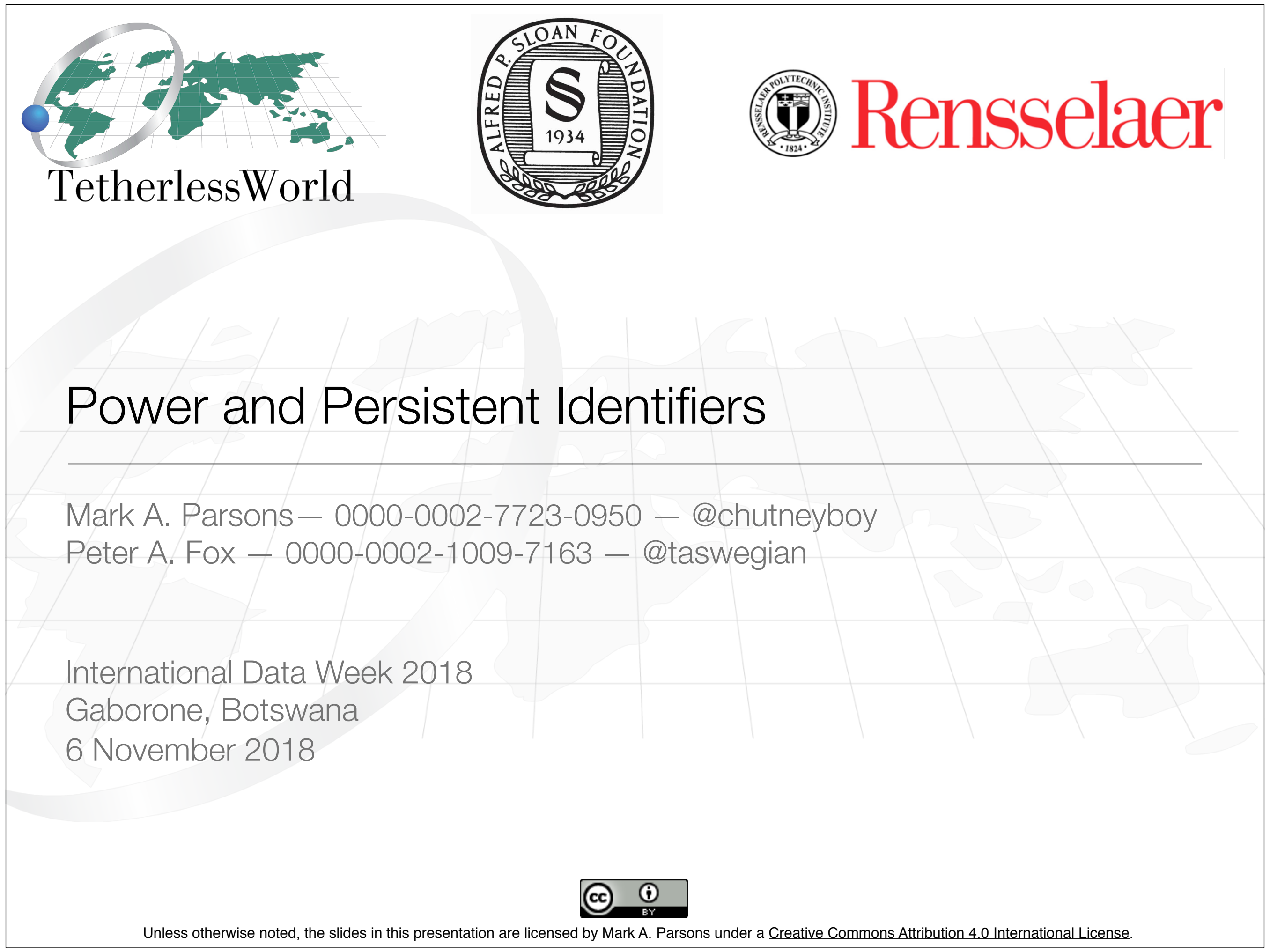


TetherlessWorld

ALFRED P. SLOAN FOUNDATION 1934

Rensselaer

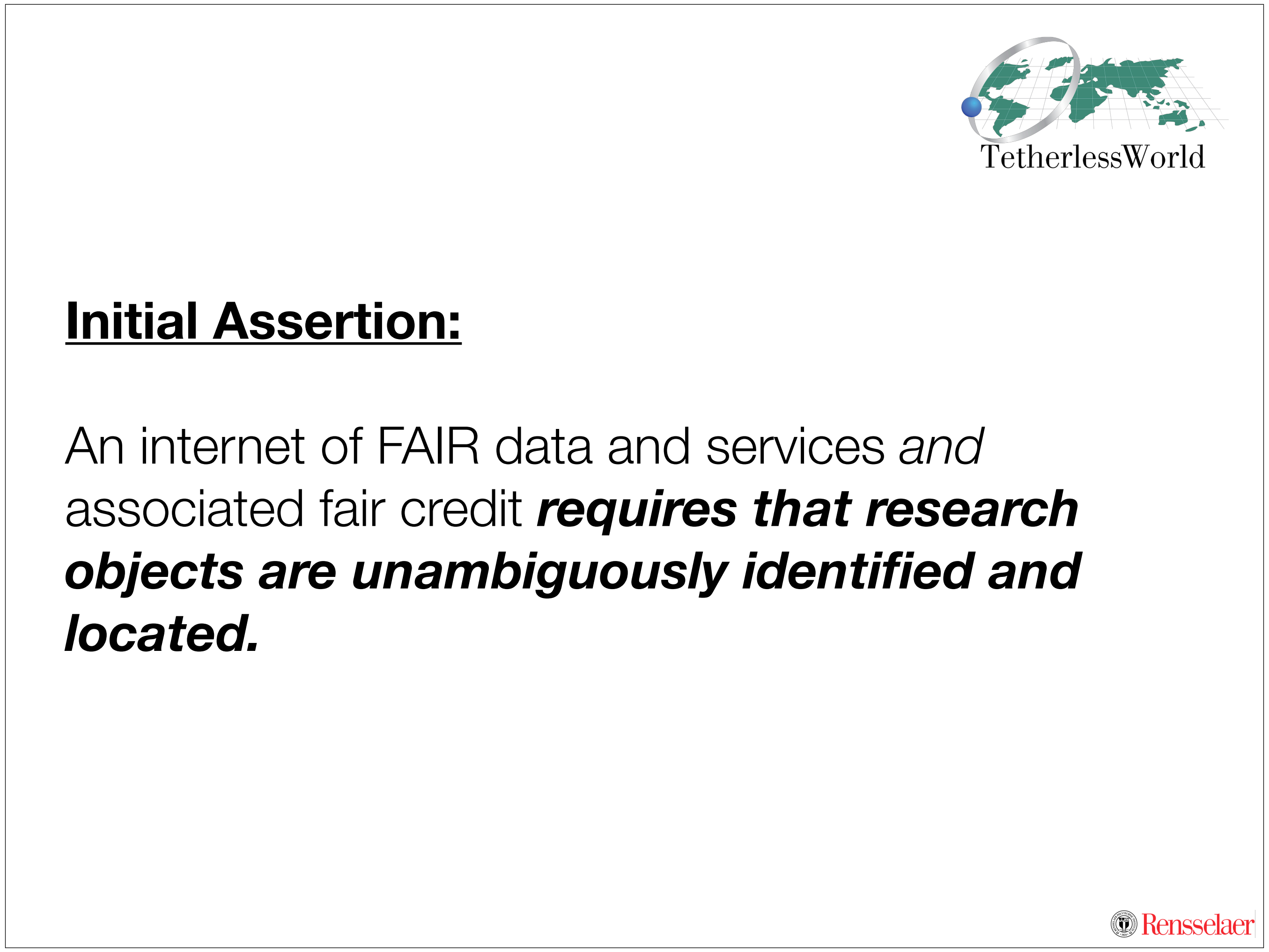
Power and Persistent Identifiers

Mark A. Parsons— 0000-0002-7723-0950 — @chutneyboy
Peter A. Fox — 0000-0002-1009-7163 — @taswegian

International Data Week 2018
Gaborone, Botswana
6 November 2018

I will argue that PIDs (handles, DOIs, pub med IDs, etc) are essential, but they create a power dynamic that must consciously be considered in their implementation.

This is research paper, and a contemplative essay. I provide no great answers. I am more concerned in how we frame our thinking. How we define what is important and how those definitions are made by whom. My goal is really just to get you to think, but I will use some real world examples of implementation of some formal RDA Recommendations.



TetherlessWorld

Initial Assertion:

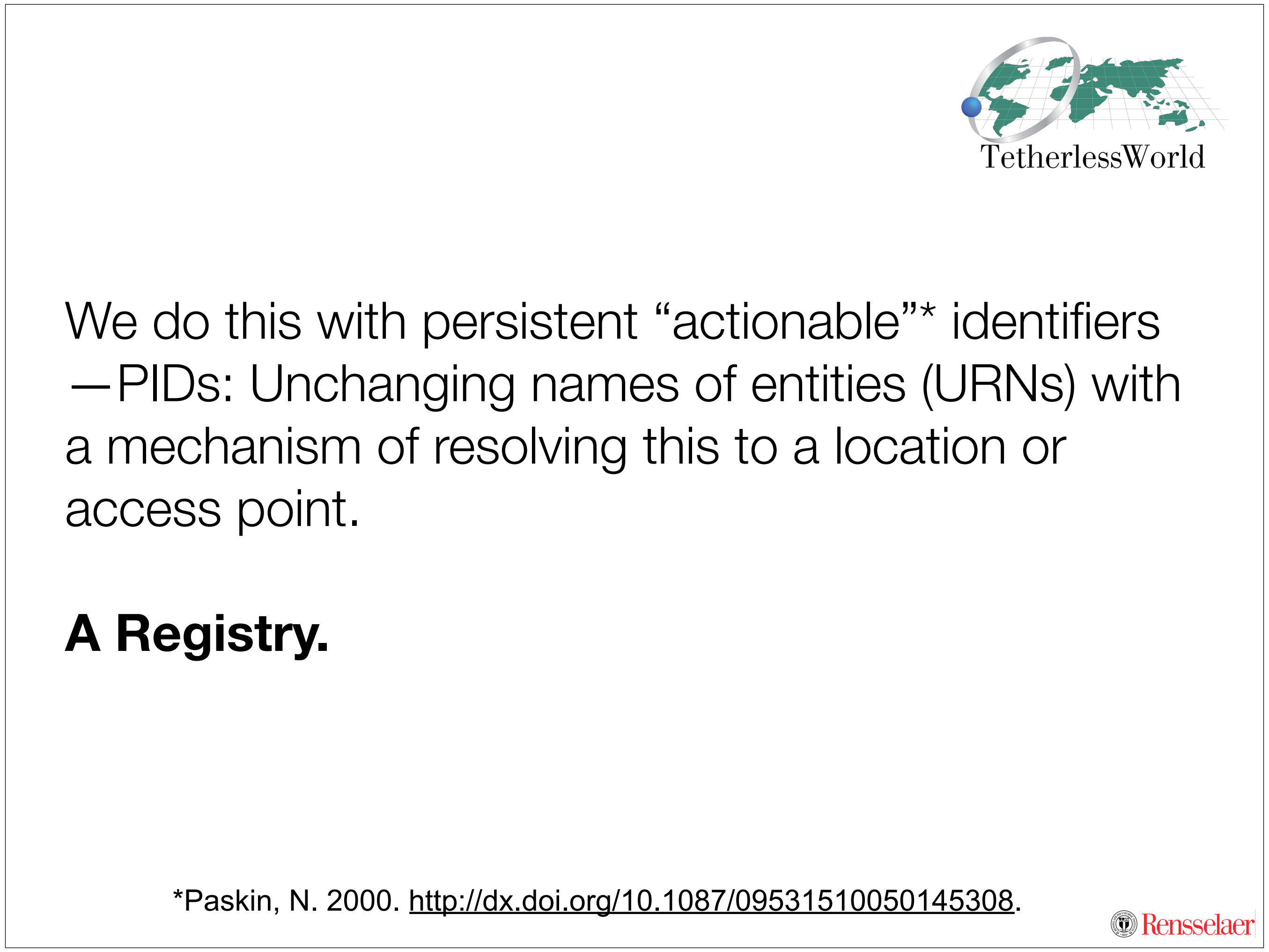
An internet of FAIR data and services *and* associated fair credit ***requires that research objects are unambiguously identified and located.***

Rensselaer

I begin with a basic assertion built around the topics of this session.

There’s a lot of work on this. It has been somewhat of an embedded principle for RDA. In short, you cannot find, access, interoperate or reuse something unless you know what and where it is. This is especially true if you are a machine.

Also, we cannot credit the creation of an object without being sure what it is and where it came from



We do this with persistent “actionable”* identifiers—PIDs: Unchanging names of entities (URNs) with a mechanism of resolving this to a location or access point.

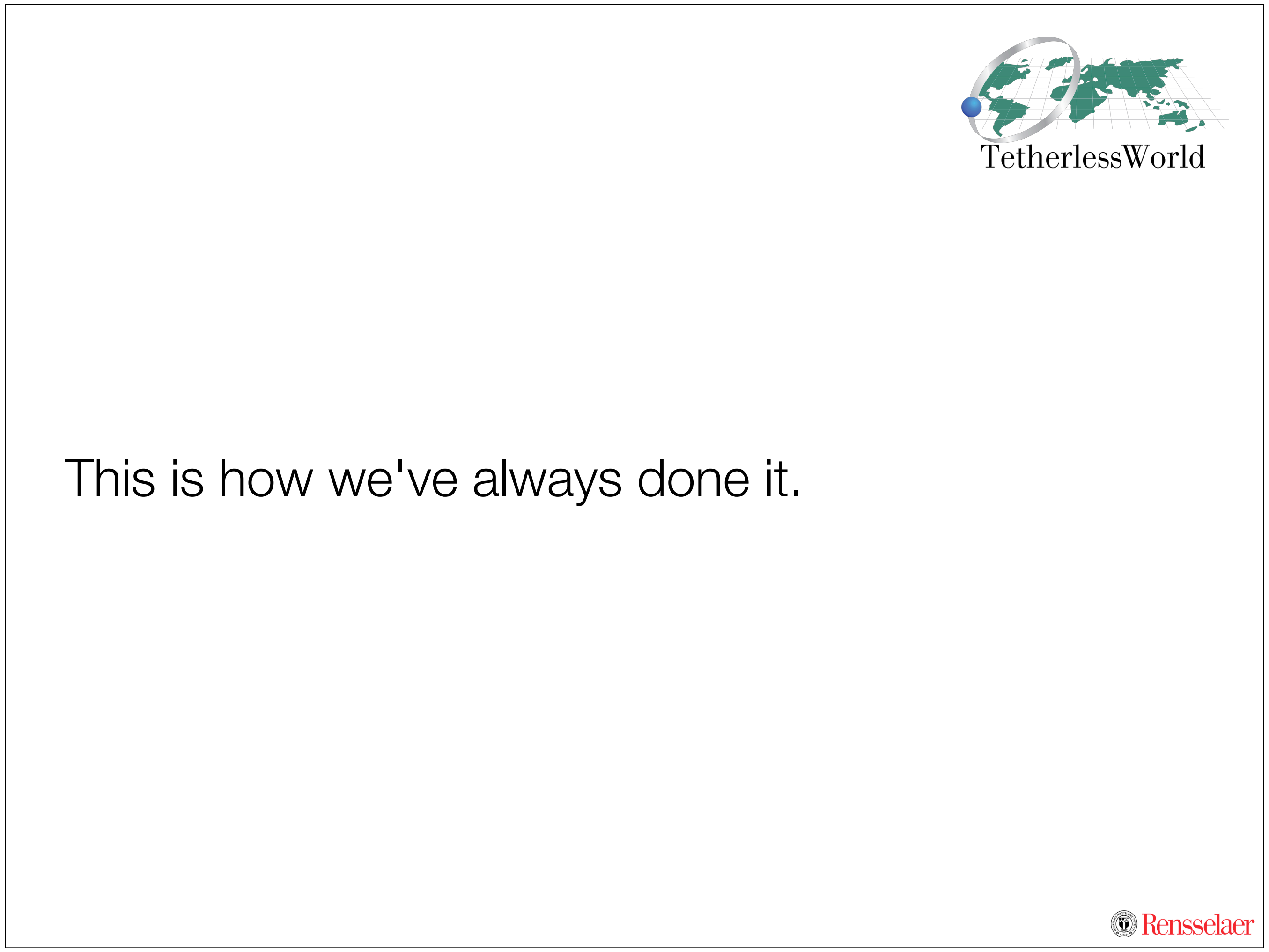
A Registry.

*Paskin, N. 2000. http://dx.doi.org/10.1087/09531510050145308.

Blockchain hasn’t changed this
Hash-based identifiers in peer to peer systems haven’t changed this.

The two functions—name and location—must still be addressed, and location in particular has a a fundamental, sustained, institutional component independent of technology.

But anyway, registries



TetherlessWorld

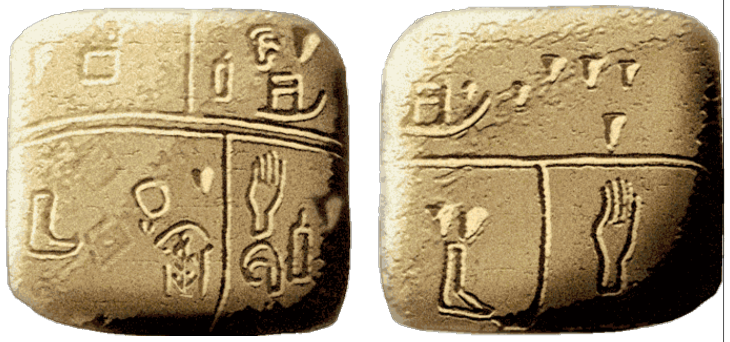
This is how we've always done it.

Rensselaer

One might even consider it a basis of modern civilization

Kish Tablet
3500 BCE

A proto-writing system used in Kish, Mesopotania. Dated from 3500 BCE. A **method of keeping accounts** by engraving. (Ashmolean Museum, Oxford, UK.)



More than 5000 years ago people started recording assets.

From the mesopotamic city of Kish (Iraq), dated from 3500 BC. Probably, it is the earliest known evidence of writing, and contains pictographs of heads, feet, hands, numbers and threshing-boards. Department of Antiquities, Ashmolean Museum, Oxford (United Kingdom).

Domesday Book 1086

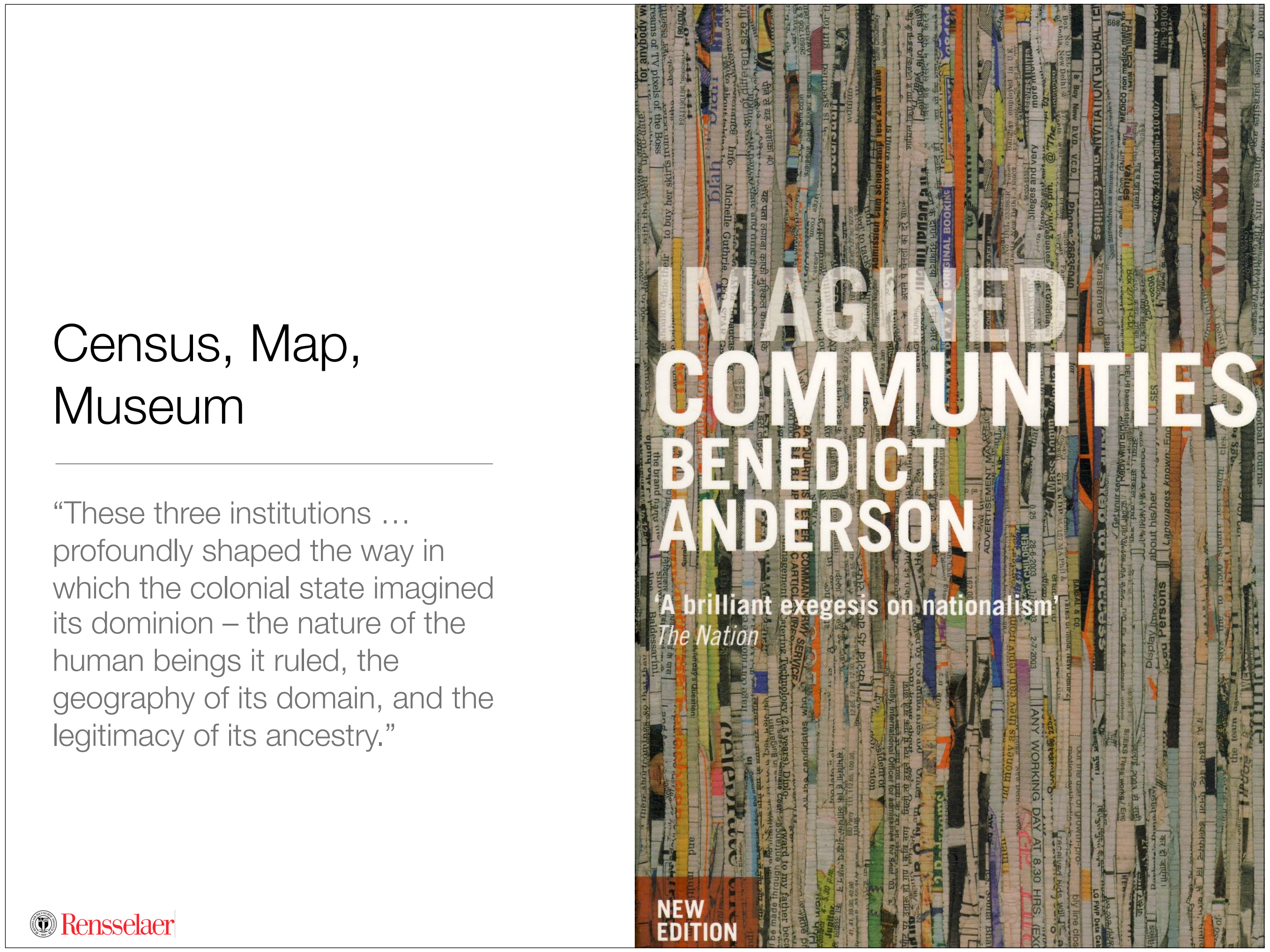
“there was no single hide nor a yard of land, nor indeed one ox nor one cow nor one pig which was left out”

The Domesday Book 1086
and William the Bastard/Conquerer
AncientPages.com

A more modern example that illustrates the power dynamic of a registry. [tell story]

changed the course of history and much of the english language.

And indeed we might argue registries are central to power, especially political power…



Census, Map, Museum

"These three institutions … profoundly shaped the way in which the colonial state imagined its dominion – the nature of the human beings it ruled, the geography of its domain, and the legitimacy of its ancestry."

This is illustrated in one of the most cited books in social sciences, in chapter 5. CMM, which argued that […]

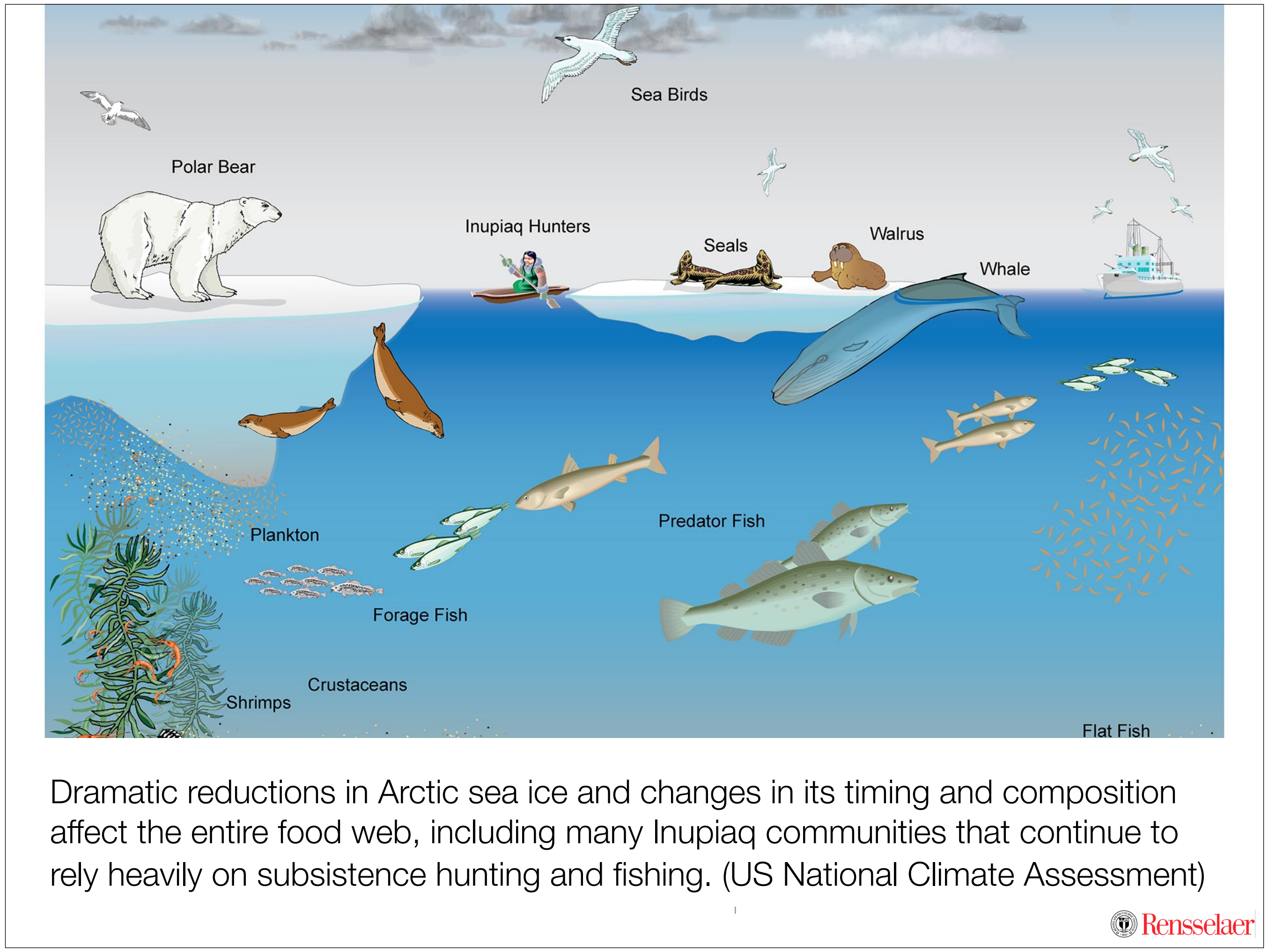
I note, that these are all, in essence, registries. Indeed a central point of Anderson's is how standardization contributed to the imaginings of national identity.

I want to emphasize that I am NOT dwelling on colonialism but simply noting how powers that be define what is considered to be important.

Let me provide an example from the other end of the world.

A map anticipated spatial reality, not vice versa. In other words, a map was a model for, rather than a model of, what it purported to represent. . . .

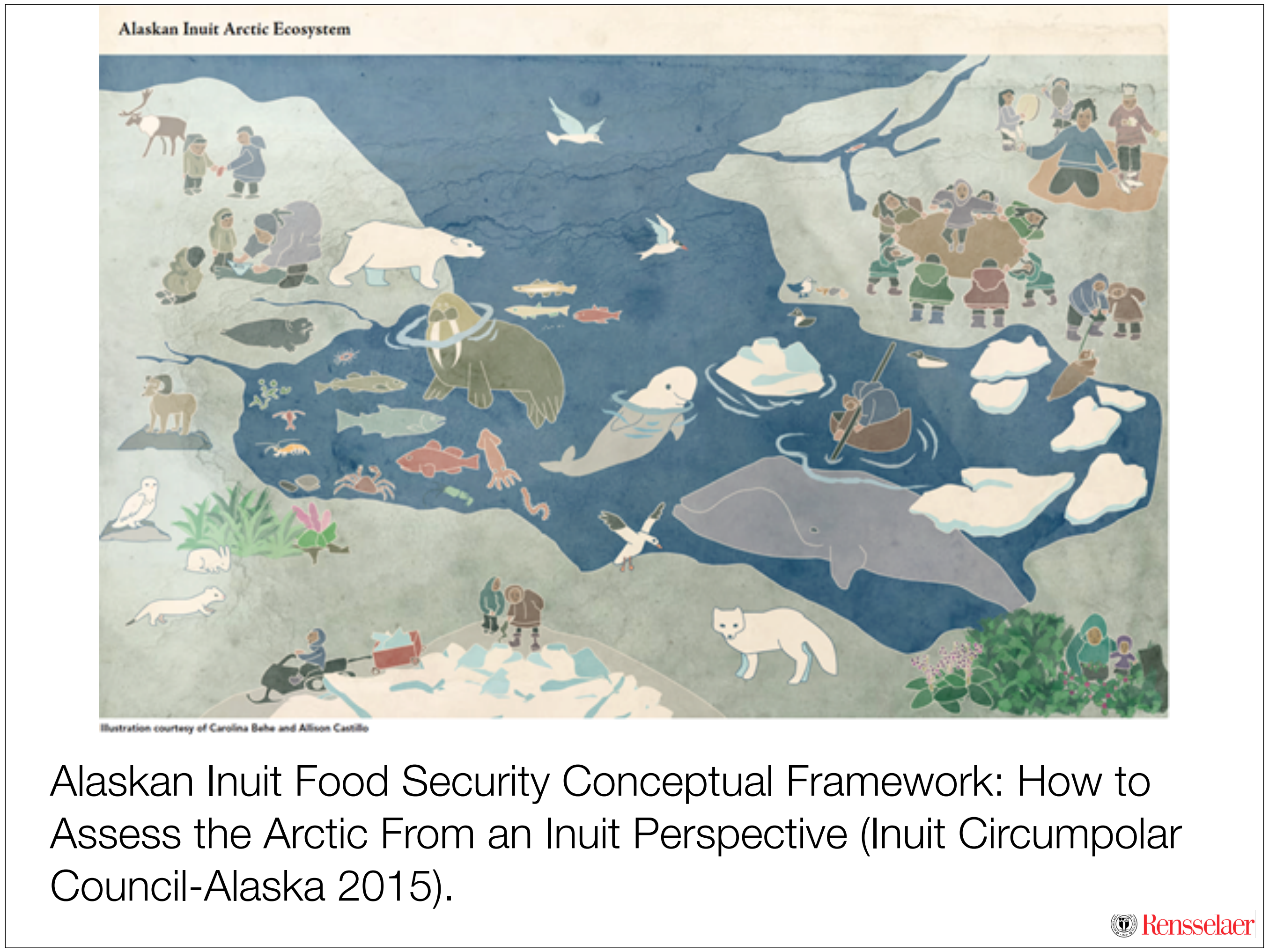
Anderson, Benedict. Imagined Communities: Reflections on the Origin and Spread of Nationalism (p. 177). Verso Books. Kindle Edition.



Dramatic reductions in Arctic sea ice and changes in its timing and composition affect the entire food web, including many Inupiaq communities that continue to rely heavily on subsistence hunting and fishing. (US National Climate Assessment)

The US climate assessment describes how Dramatic reductions in Arctic sea ice and changes in its timing and composition affect the entire food web, including many Inupiaq communities — the Indigenous people of northern AK—They provide this cross-section style figure to illustrate the point.

https://nca2014.globalchange.gov/report/sectors/indigenous-peoples/graphics/arctic-marine-food-web



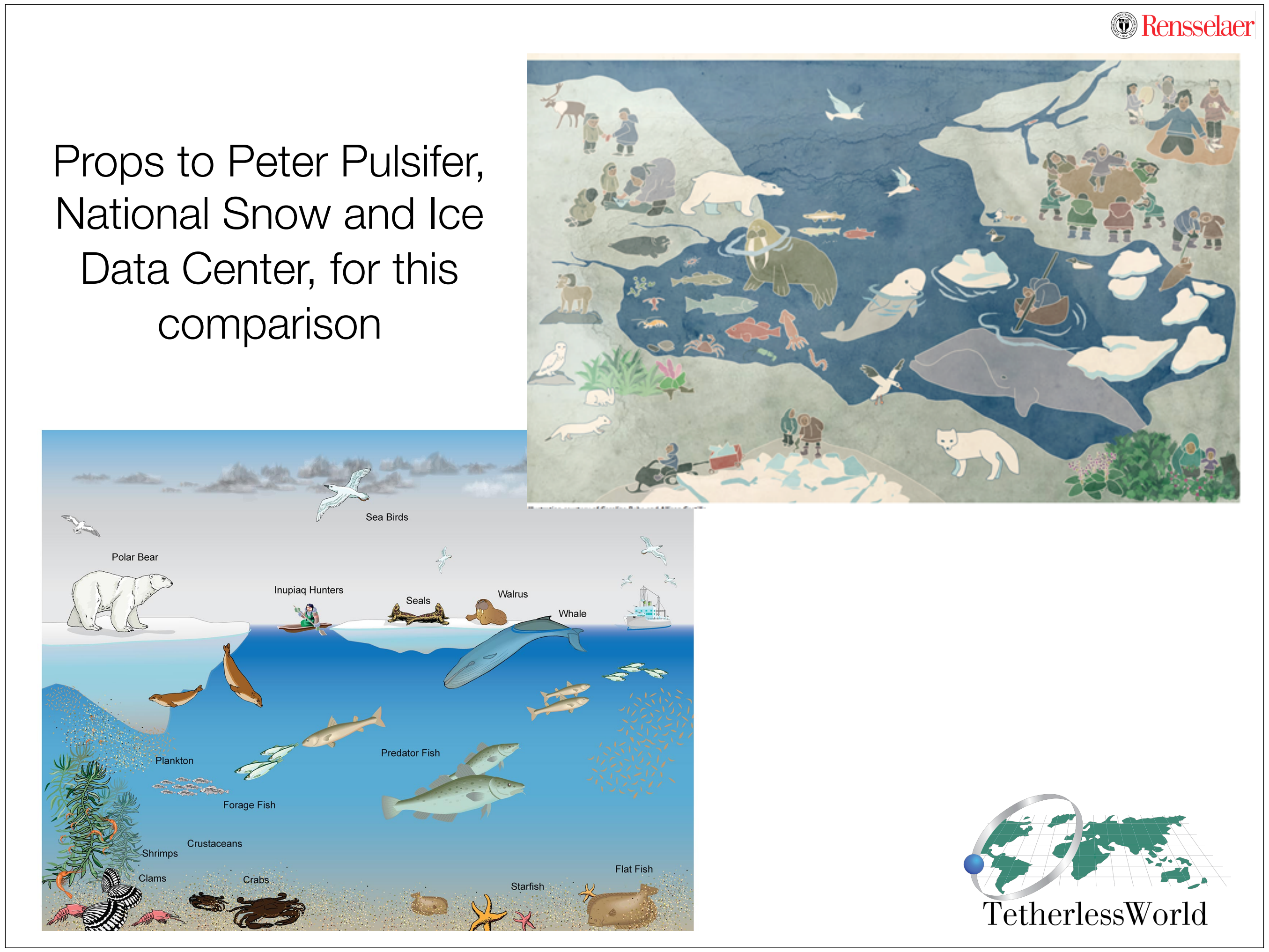
Alaskan Inuit Arctic Ecosystem

Illustration courtesy of Carolina Behe and Allison Castillo

Alaskan Inuit Food Security Conceptual Framework: How to Assess the Arctic From an Inuit Perspective (Inuit Circumpolar Council-Alaska 2015).

Here’s another view from the Inuit Circumpolar Council, representatives of those local people. It’s a birds-eye view with some depth. More 3D and artistic, and more inclusive, especially in terms of local human activity. (Although the commercial ship is excluded)

Inuit Circumpolar Council-Alaska 2015. Alaskan Inuit Food Security Conceptual Framework: How to Assess the Arctic From an Inuit Perspective: Summary Report and Recommendations Report. Anchorage, AK.



Props to Peter Pulsifer, National Snow and Ice Data Center, for this comparison

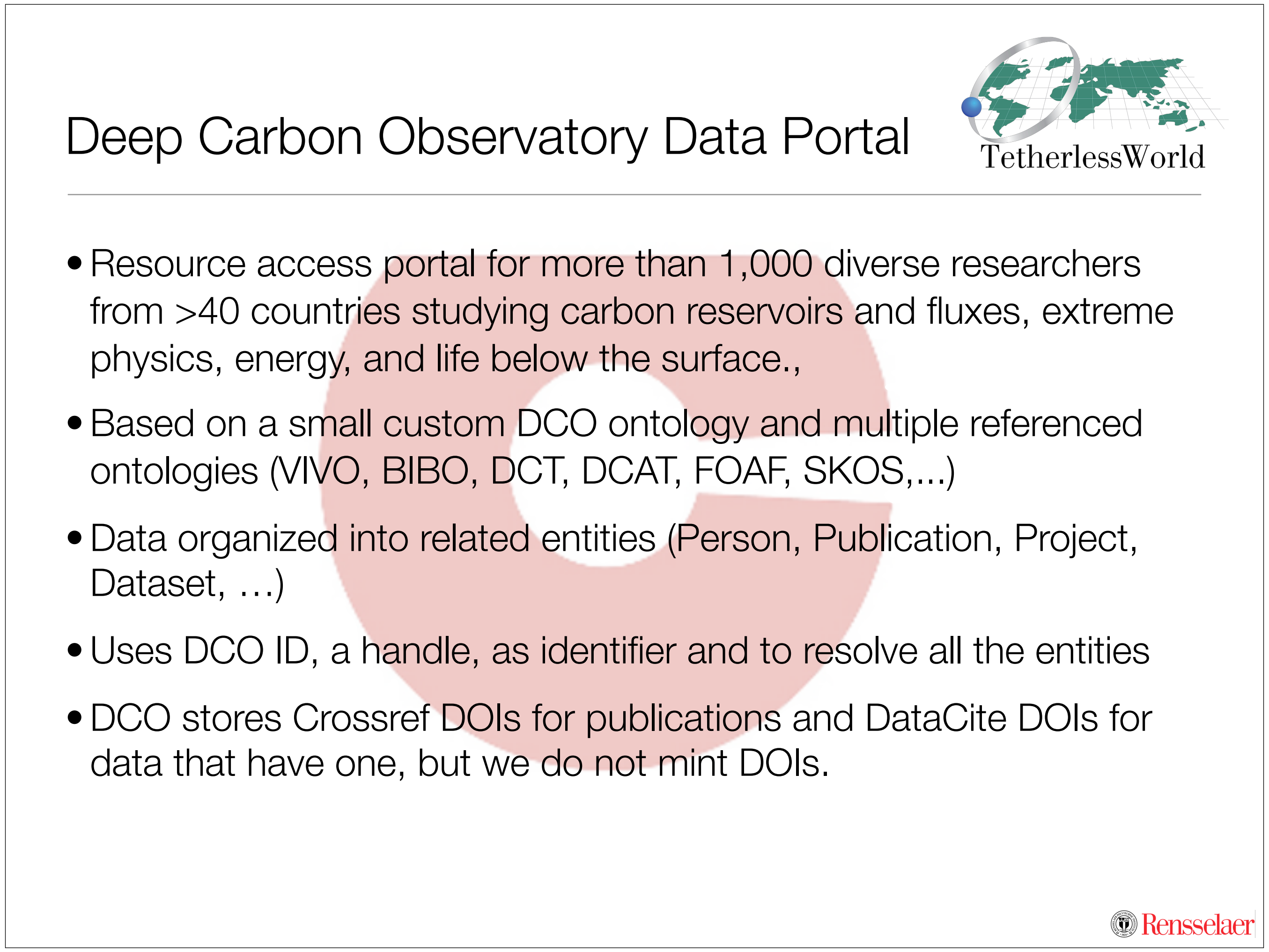
Maybe it helps to see them next to each other. They have a lot in common, but they clearly look at things differently. So different that they would construct quite different resource identification schemes and data systems.

It is not a question of right or wrong, but rather who defines what is important for a context.

I also note that often cultures based in an oral tradition have more descriptive and verb-based place names. Suggesting a focus on flows and connections can also identify and provide a level of persistence.

So given that broad context, let me now turn to some specific issues around PIDs and data sharing and scholarly communication, where much of the conversation is occurring.

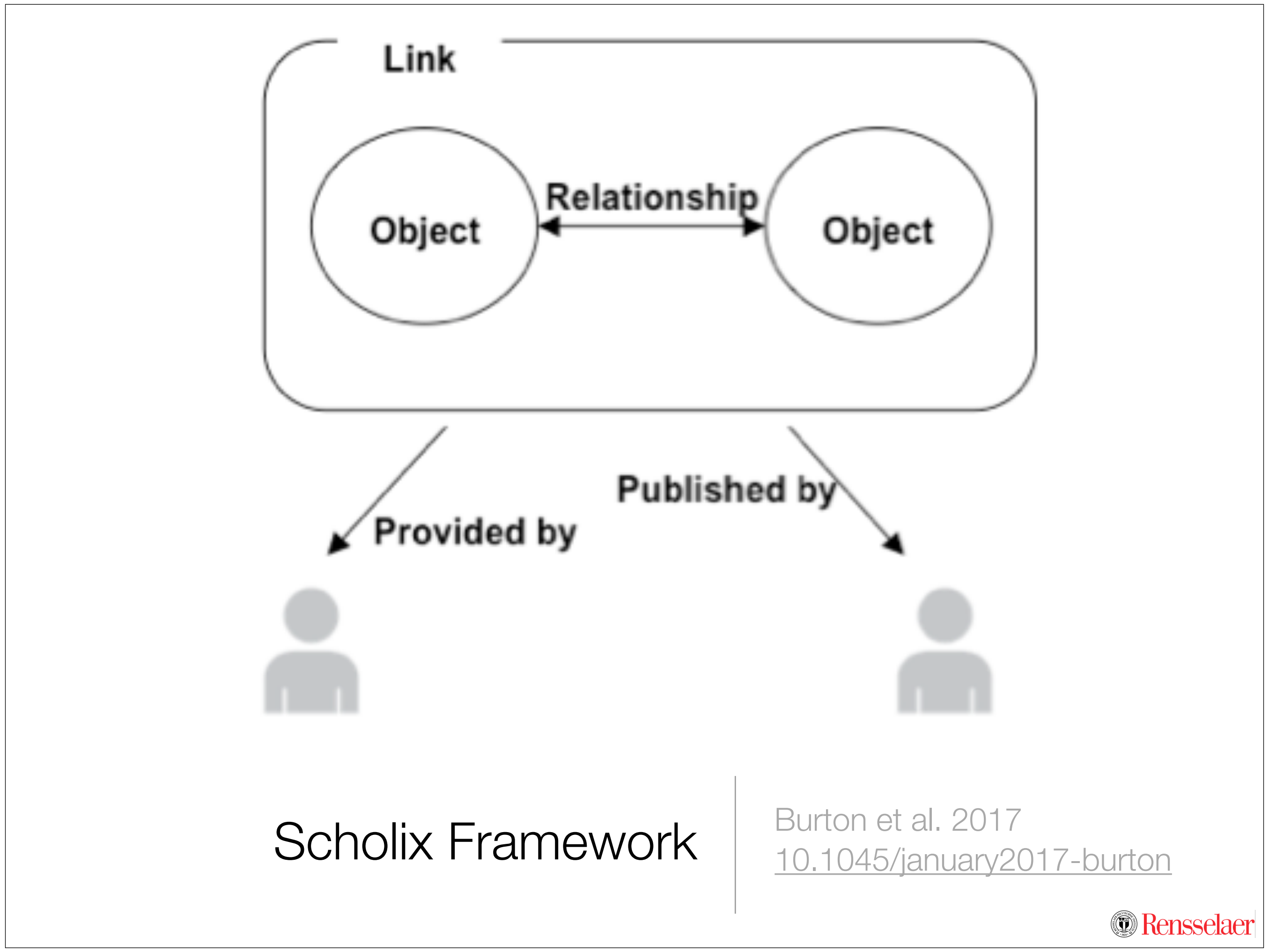
I must note in this context that power doesn't mean nefarious intent, but it can mislead because of how it can frame critical questions based on assumptions about what things are and how they are used.



Deep Carbon Observatory Data Portal

- Resource access portal for more than 1,000 diverse researchers from >40 countries studying carbon reservoirs and fluxes, extreme physics, energy, and life below the surface.,
- Based on a small custom DCO ontology and multiple referenced ontologies (VIVO, BIBO, DCT, DCAT, FOAF, SKOS,...)
- Data organized into related entities (Person, Publication, Project, Dataset, …)
- Uses DCO ID, a handle, as identifier and to resolve all the entities
- DCO stores Crossref DOIs for publications and DataCite DOIs for data that have one, but we do not mint DOIs.

DCO has been working to adopt a number of RDA Recommendations. One of these is Scholix a mechanism to capture links between different objects.



Link

Object

Relationship

Object

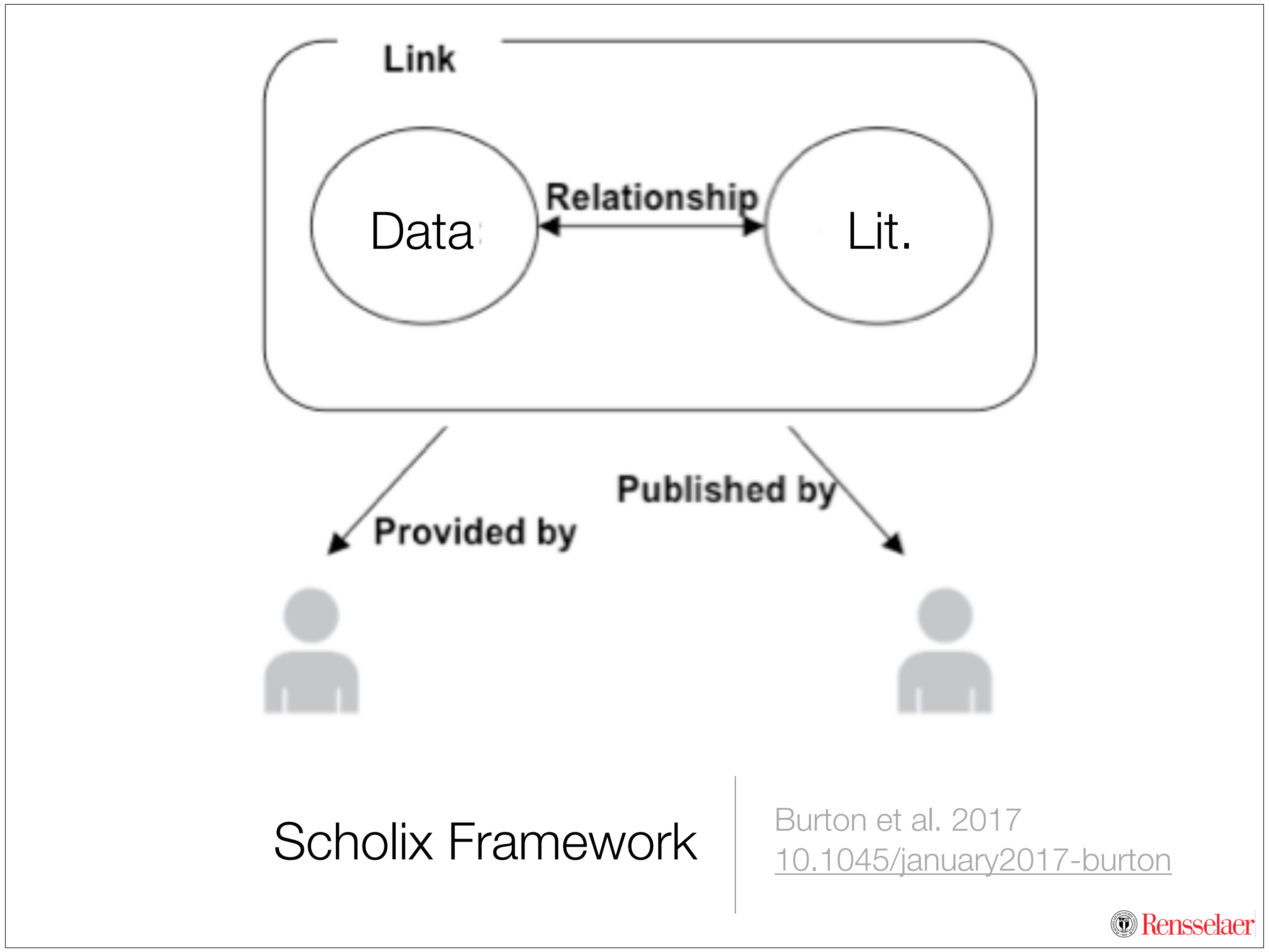
Provided by

Published by

Scholix Framework

Burton et al. 2017
10.1045/january2017-burton

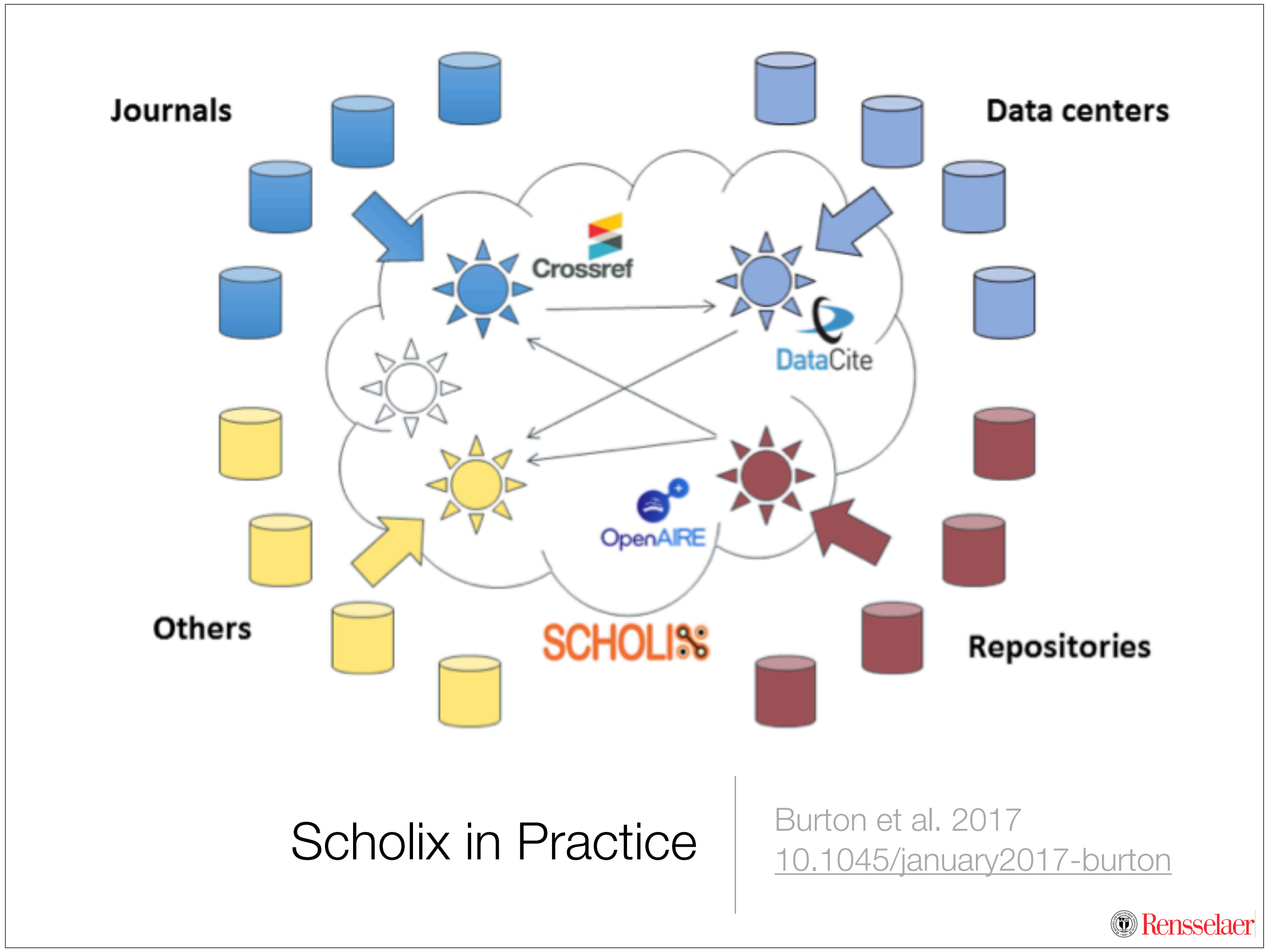
Scholix is based on a fundamental relationship, but they wisely choose to simplify the problem and they initially focus on on just data and literature. And they make some
assumptions based on the traditional publishing scheme with traditional assets and roles.



Scholix Framework

Burton et al. 2017
10.1045/january2017-burton

Scholix is based on a fundamental relationship, but they wisely choose to simplify the problem and they initially focus on on just data and literature. And they make some
assumptions based on the traditional publishing scheme with traditional assets and roles.



Scholix in Practice

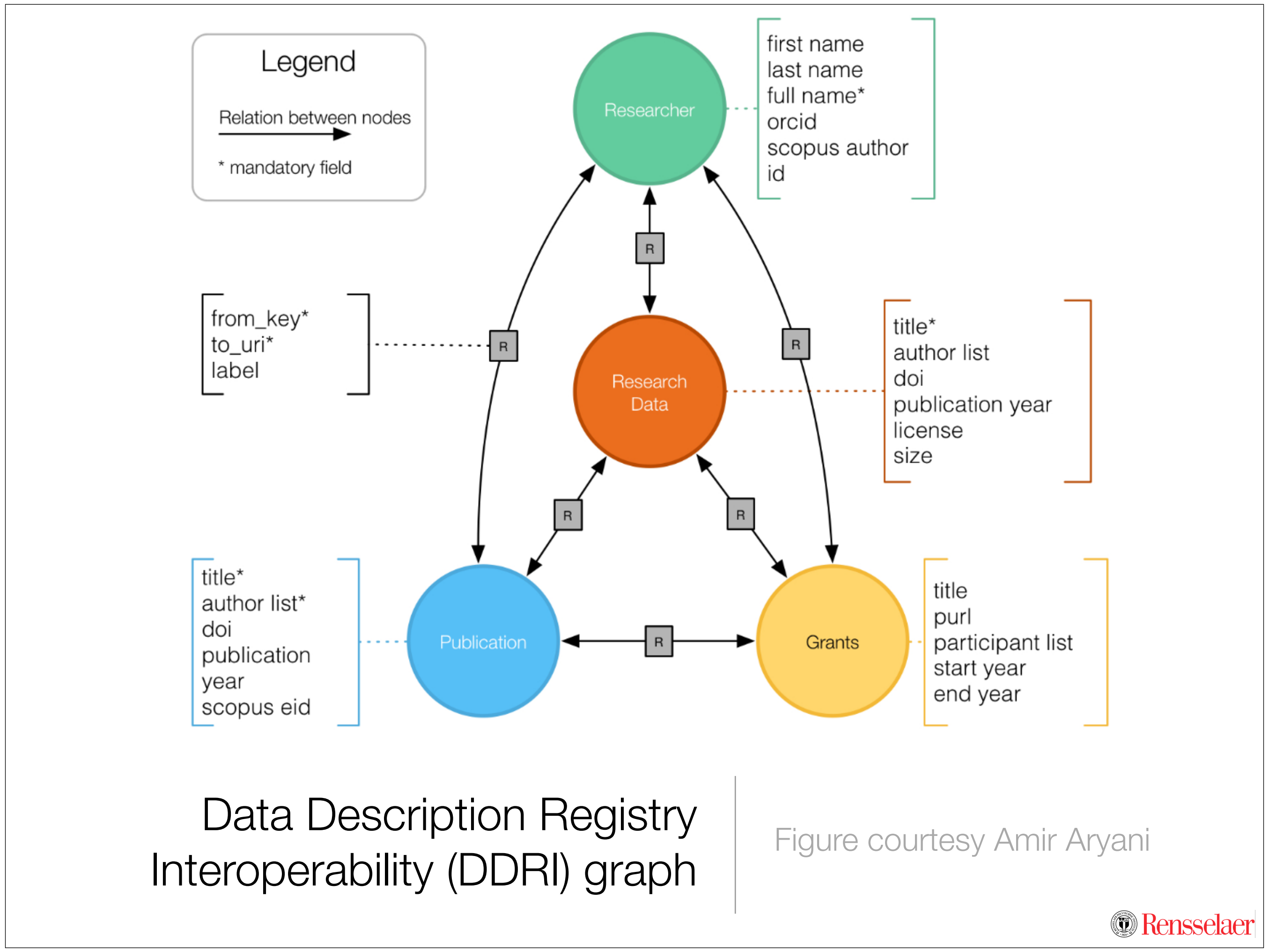
Burton et al. 2017
10.1045/january2017-burton

This figure comes from a paper describing the scheme. I quote: “The Scholix framework rests on a multi-hub network that collects and aggregates information about data-literature links. Each hub focuses on a particular community. Communities are for, e.g., the journal publishers, the data centers, etc. The hubs aggregate information from communities. Of particular interest are "natural" hubs, i.e. existing infrastructures that focus on community information aggregation.” i.e the powers that be.

DCO had agreed to adopt Scholix but we weren’t sure what that meant.

Be a hub and potentially take on new responsibilities?

Change our workflow and systems to feed a hub? As you recall, we don’t use DOIs, which this system largely assumes.



Legend

Relation between nodes

* mandatory field

Researcher

first name
last name
full name*
orcid
scopus author id

from_key*
to_uri*
label

R

Research Data

title*
author list
doi
publication year
license
size

Publication

title*
author list*
doi
publication year
scopus eid

Grants

title
purl
participant list
start year
end year

Data Description Registry Interoperability (DDRI) graph

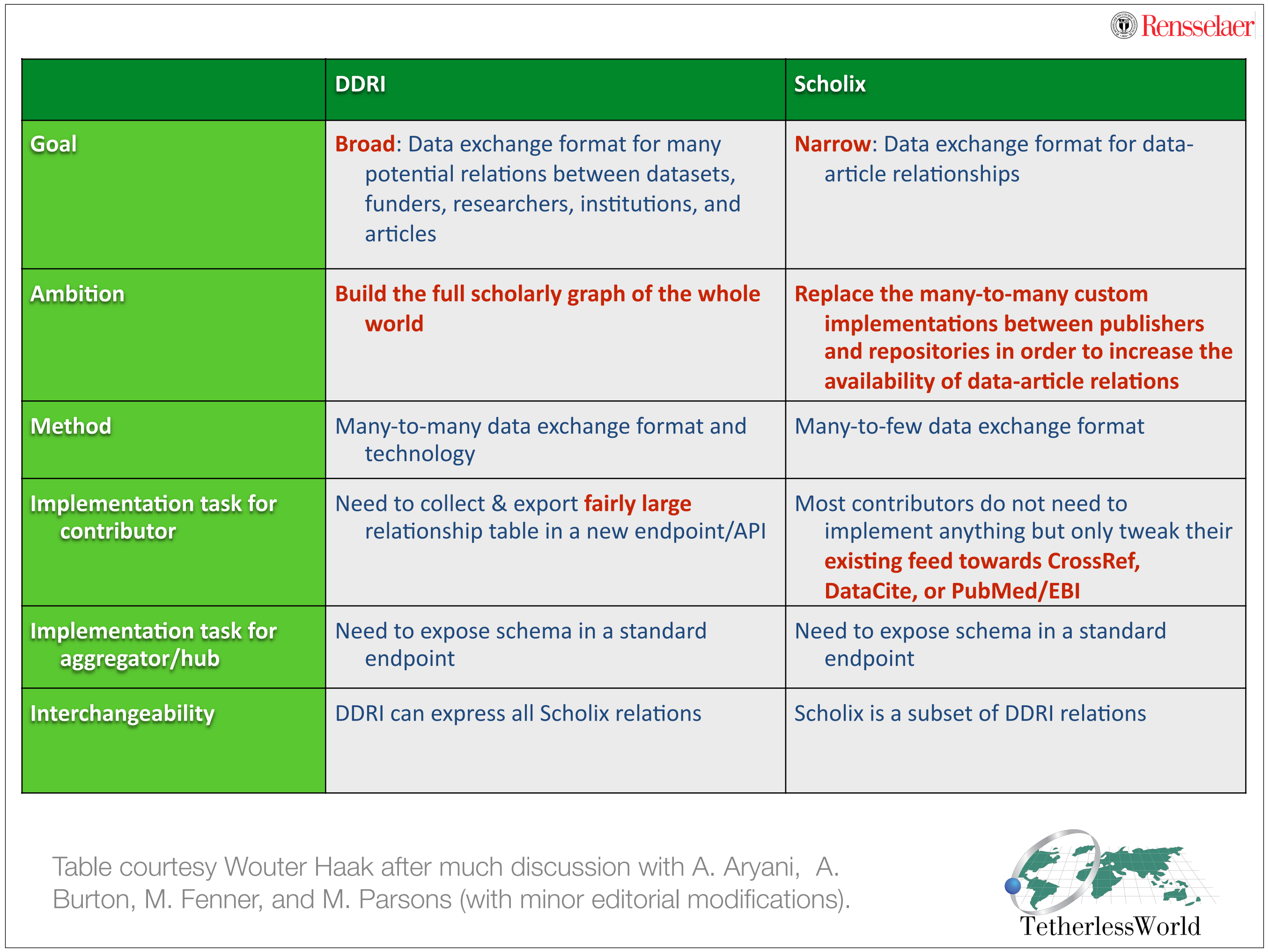
Figure courtesy Amir Aryani

Rensselaer

Then we started looking at another RDA Recommendation — DDRI

Here the research objects are defined but include more complex relationships.
The schema also explicitly introduces the researcher/author/creator/**human** who is only implicit in Scholix.

AND they mapped perfectly to the ontologies we were using! (notably VIVO)



Rensselaer

	DDRI	Scholix
Goal	**Broad**: Data exchange format for many potential relations between datasets, funders, researchers, institutions, and articles	**Narrow**: Data exchange format for data-article relationships
Ambition	**Build the full scholarly graph of the whole world**	**Replace the many-to-many custom implementations between publishers and repositories in order to increase the availability of data-article relations**
Method	Many-to-many data exchange format and technology	Many-to-few data exchange format
Implementation task for contributor	Need to collect & export **fairly large** relationship table in a new endpoint/API	Most contributors do not need to implement anything but only tweak their **existing feed towards CrossRef, DataCite, or PubMed/EBI**
Implementation task for aggregator/hub	Need to expose schema in a standard endpoint	Need to expose schema in a standard endpoint
Interchangeability	DDRI can express all Scholix relations	Scholix is a subset of DDRI relations

Table courtesy Wouter Haak after much discussion with A. Aryani, A. Burton, M. Fenner, and M. Parsons (with minor editorial modifications).

TetherlessWorld

DDRI and Scholix people are working together and they are interrelated, but there are clear differences.

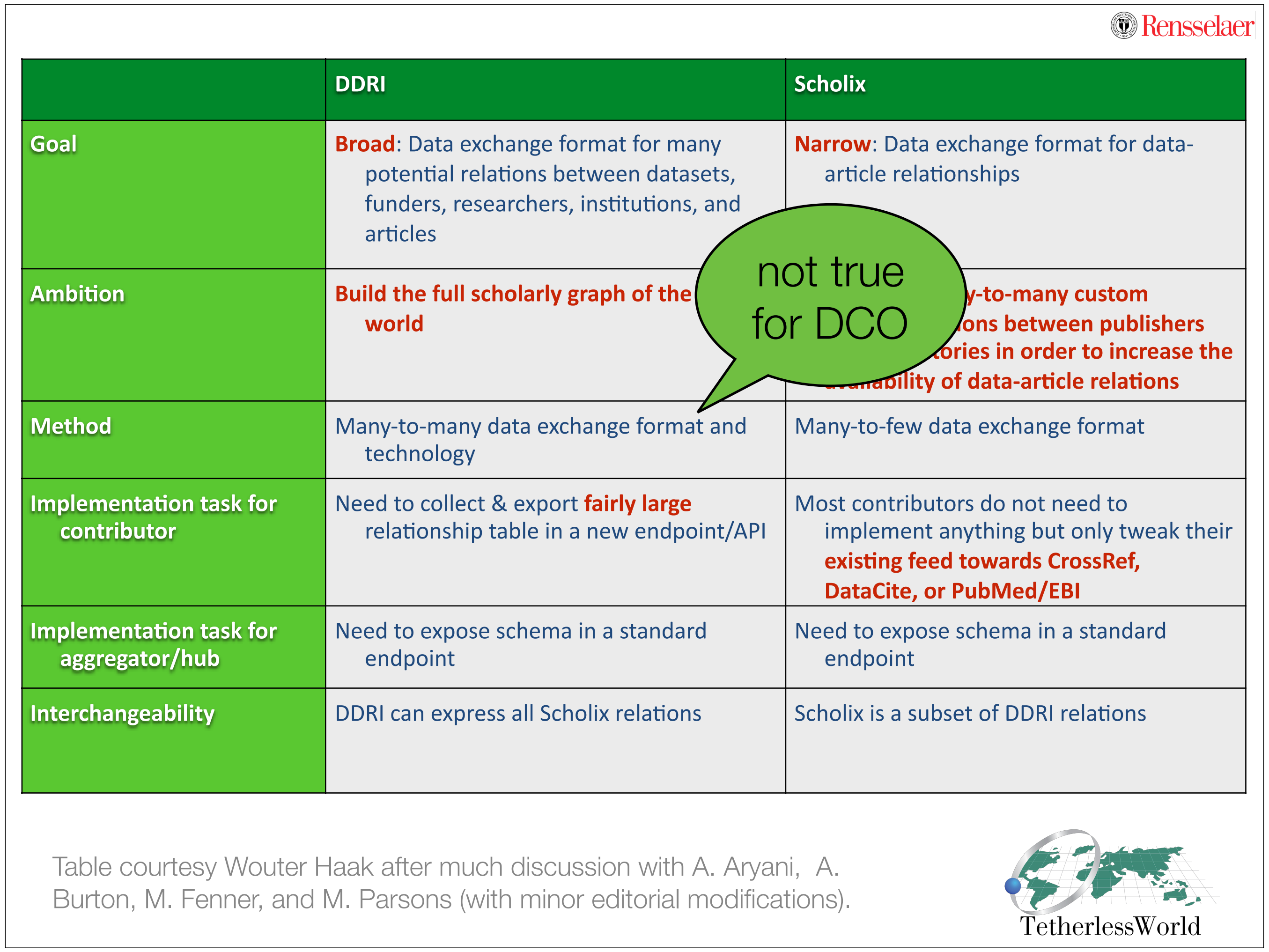
It took me a number of conversations to sort it out, and this table nicely created by Wouter Haak summarizes:

[hit high points then hit ballon and note the power brokers and the assumptions of how easy things are based on that and how for DCO DDRI was much easier]

Ultimately the DCO team concluded that we should adopt the DDRI recommendations due to their greater breadth, depth, and flexibility. And then work to ensure our records could be consumed by both. This maintained local control and ceded no authority while still ensuring interoperability.

Note, I’m not trying to portray this a power play or to dis Scholix, which has made great advances, but simply to allow more local control, more self determination.

Let me now give a second related example.



Rensselaer

	DDRI	Scholix
Goal	**Broad**: Data exchange format for many potential relations between datasets, funders, researchers, institutions, and articles	**Narrow**: Data exchange format for data-article relationships
Ambition	**Build the full scholarly graph of the world**	**y-to-many custom ons between publishers tories in order to increase the ability of data-article relations**
Method	Many-to-many data exchange format and technology	Many-to-few data exchange format
Implementation task for contributor	Need to collect & export **fairly large** relationship table in a new endpoint/API	Most contributors do not need to implement anything but only tweak their **existing feed towards CrossRef, DataCite, or PubMed/EBI**
Implementation task for aggregator/hub	Need to expose schema in a standard endpoint	Need to expose schema in a standard endpoint
Interchangeability	DDRI can express all Scholix relations	Scholix is a subset of DDRI relations

not true for DCO

Table courtesy Wouter Haak after much discussion with A. Aryani, A. Burton, M. Fenner, and M. Parsons (with minor editorial modifications).

TetherlessWorld

DDRI and Scholix people are working together and they are interrelated, but there are clear differences.

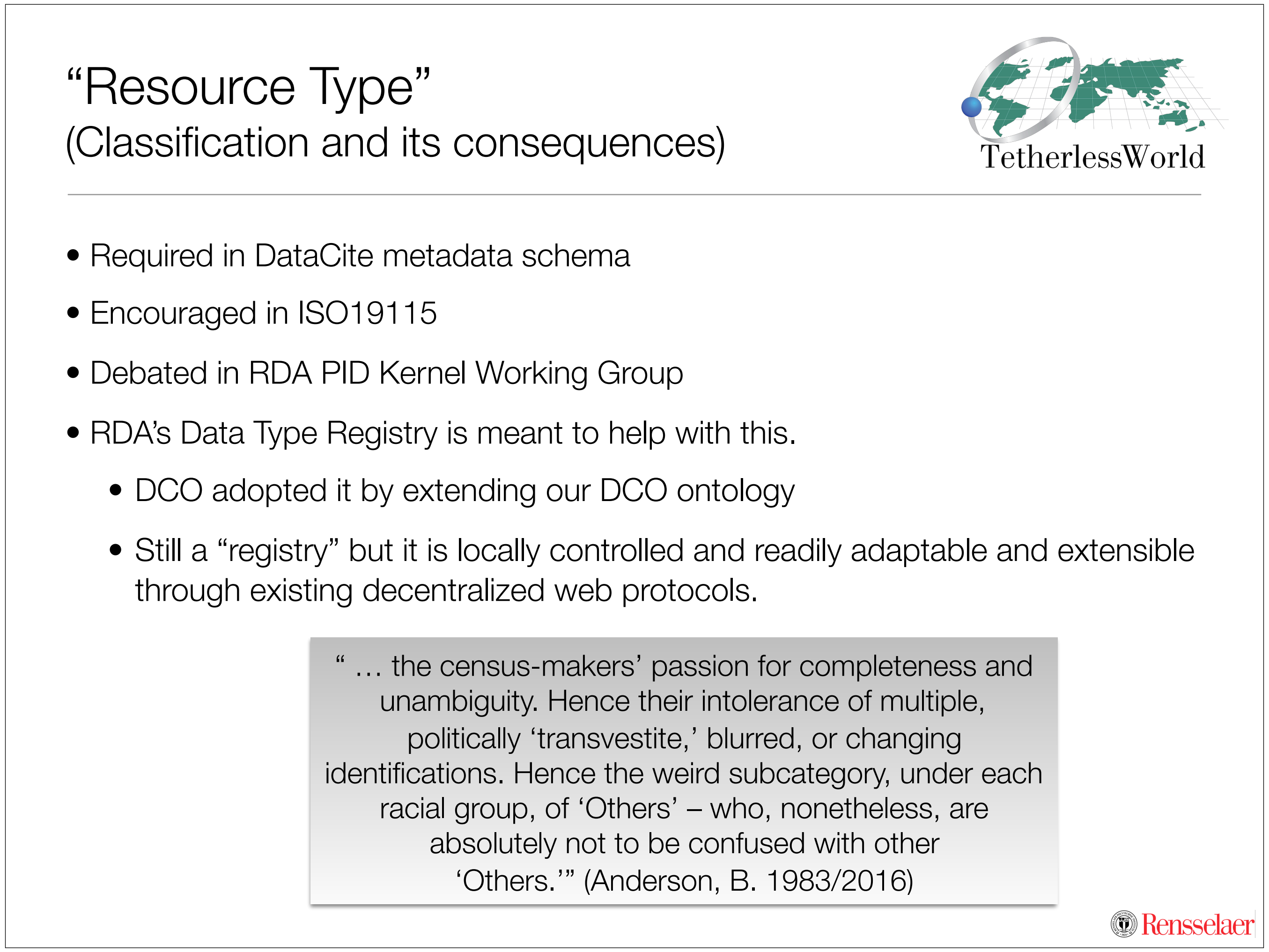
It took me a number of conversations to sort it out, and this table nicely created by Wouter Haak summarizes:

[hit high points then hit ballon and note the power brokers and the assumptions of how easy things are based on that and how for DCO DDRI was much easier]

Ultimately the DCO team concluded that we should adopt the DDRI recommendations due to their greater breadth, depth, and flexibility. And then work to ensure our records could be consumed by both. This maintained local control and ceded no authority while still ensuring interoperability.

Note, I'm not trying to portray this a power play or to dis Scholix, which has made great advances, but simply to allow more local control, more self determination.

Let me now give a second related example.



“Resource Type”
(Classification and its consequences)

TetherlessWorld

- Required in DataCite metadata schema
- Encouraged in ISO19115
- Debated in RDA PID Kernel Working Group
- RDA’s Data Type Registry is meant to help with this.
    - DCO adopted it by extending our DCO ontology
    - Still a “registry” but it is locally controlled and readily adaptable and extensible through existing decentralized web protocols.

> “ … the census-makers’ passion for completeness and unambiguity. Hence their intolerance of multiple, politically ‘transvestite,’ blurred, or changing identifications. Hence the weird subcategory, under each racial group, of ‘Others’ – who, nonetheless, are absolutely not to be confused with other ‘Others.’” (Anderson, B. 1983/2016)

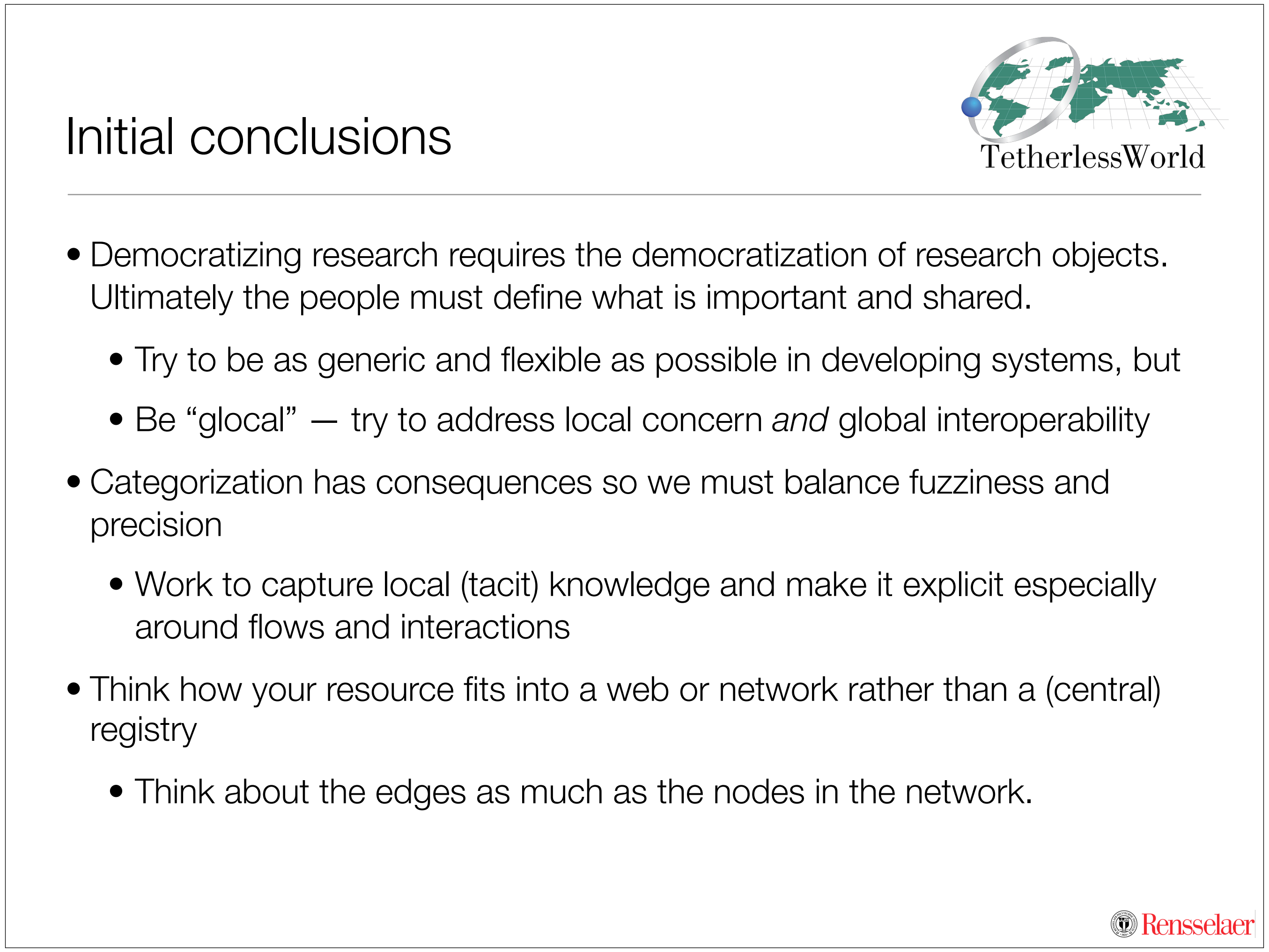
Rensselaer

Much of the tension that emerges from these systems is around typing or categorization. Categories are continuously agglomerated, disaggregated, recombined, intermixed, and reordered, but the politically powerful identity categories always lead the list.

There is a constant struggle between the ‘fuzzy and neat’ (per Lindsay Poirier), and I can only begin to go down the path of category theory, but a first step in categorizing something is identifying it.

Once you put a PID on a thing, you have already made decisions about what kind of thing it is!

Nonetheless, Resource type is required in the DataCite metadata schema and encouraged in others. RDA seeks to address this by creating a framework to “register” types. DCO adopted this Recommendation by extending our DCO ontology. It
is still a “registry” in a sense, but it is locally controlled and readily adaptable and extensible through existing decentralized web protocols.



Initial conclusions

TetherlessWorld

- Democratizing research requires the democratization of research objects. Ultimately the people must define what is important and shared.
  - Try to be as generic and flexible as possible in developing systems, but
  - Be “glocal” — try to address local concern *and* global interoperability
- Categorization has consequences so we must balance fuzziness and precision
  - Work to capture local (tacit) knowledge and make it explicit especially around flows and interactions
- Think how your resource fits into a web or network rather than a (central) registry
  - Think about the edges as much as the nodes in the network.

Rensselaer

As mentioned, my primary goal was to get y’all thinking in new ways about basic infrastructure, but I do think we can draw some initial conclusions and suggestions for repositories in particular.

[read]

Early research, and not sure how this applies to Africa exactly, but I hope it provides a useful framework of thinking.

TetherlessWorld

Rensselaer

Thank You!

Mark A. Parsons— 0000-0002-7723-0950 — @chutneyboy
Peter A. Fox — 0000-0002-1009-7163 — @taswegian