RESEARCH DATA MANAGEMENT

ΔΕΔΟΜΕΝΑ, ΣΚΥΛΛΑ, ΧΑΡΥΒΔΙΣ

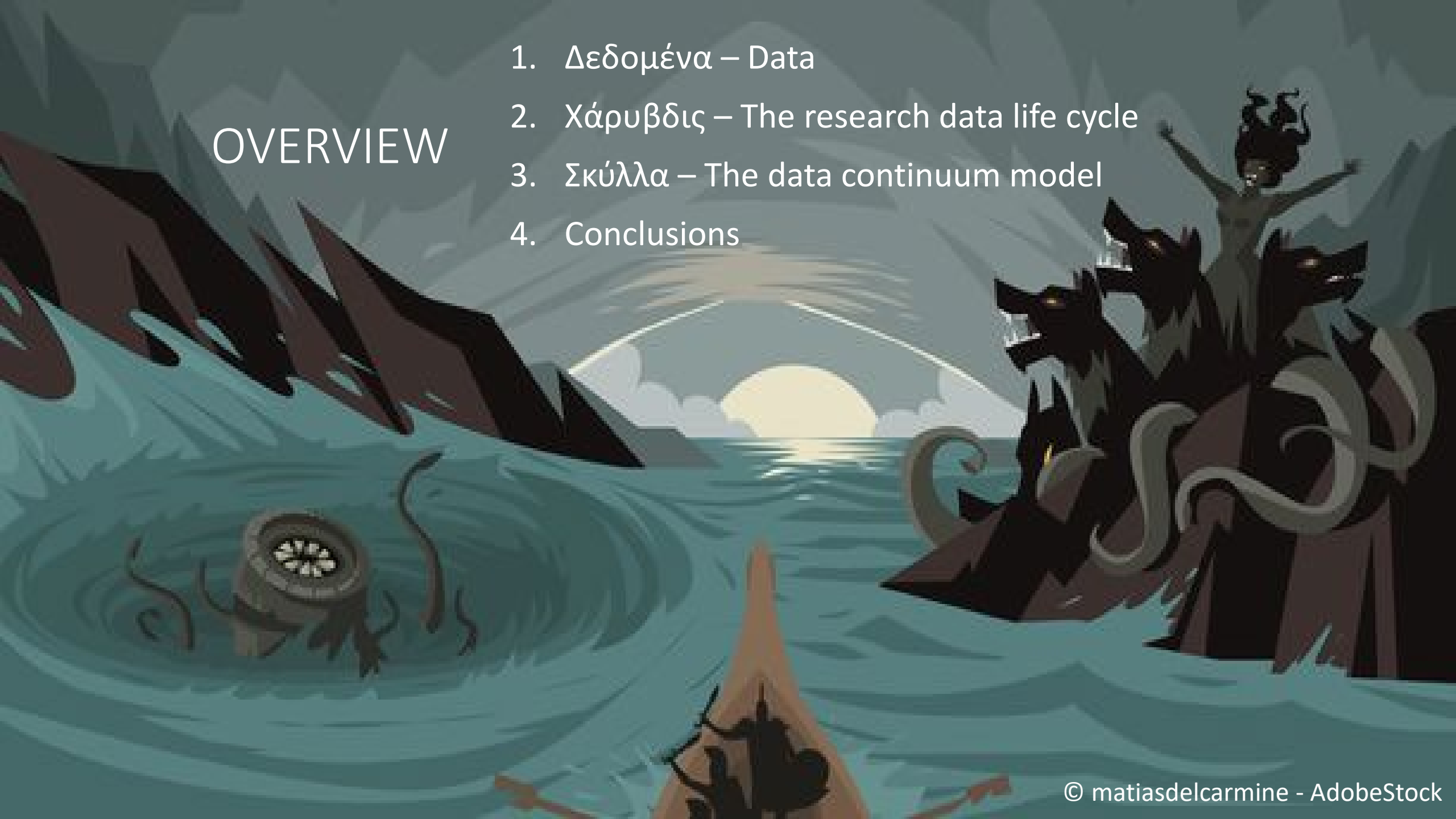
All cartoons courtesy of Jørgen Stamp, Digitalbevaring.dk. CC BY 2.5.

RENÉ SCHNEIDER

HAUTE ECOLE DE GESTION, GENEVA

OVERVIEW

1. Δεδομένα – Data
2. Χάρυβδις – The research data life cycle
3. Σκύλλα – The data continuum model
4. Conclusions



1

Δεδομένα
Data

EUKLID, 3RD CENTURY BC

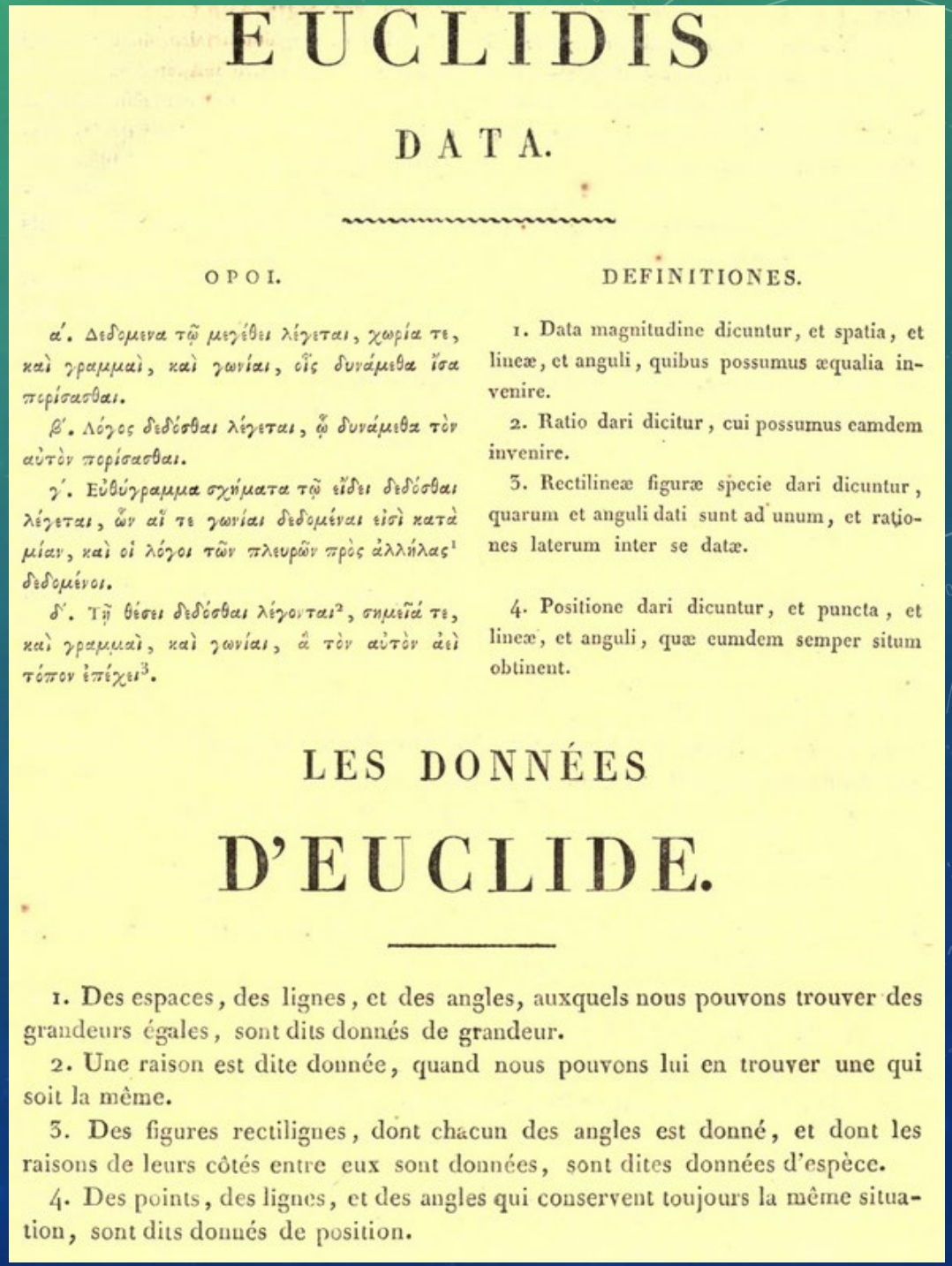
Euklid wrote first the 15 books of the «Elements» ,

still considered to be the origin of exact science.

Then he wrote «Δεδομένα», *Dedomena*, the Given, as a supplement to the «Elements».

Dedomena will then be translated to Latin as «Data» (unsurprisingly).

https://archive.org/stream/lesuvresdeuclide03eucl#page/300/mode/2up



EUCLIDIS

DATA.

OPOI.

α'. Δεδομενα τῷ μεγέθει λέγεται, χωρία τε, καὶ γραμμαὶ, καὶ γωνίαι, οἷς δυνάμεθα ἴσα περίσασθαι.

β'. Λόγος δεδόσθαι λέγεται, ᾧ δυνάμεθα τὸν αὐτὸν περίσασθαι.

γ'. Εὐθύγραμμα σχήματα τῷ εἴδει δεδόσθαι λέγεται, ὧν αἵ τε γωνίαι δεδομέναι εἰσὶ κατὰ μίαν, καὶ οἱ λόγοι τῶν πλευρῶν πρὸς ἀλλήλας[1] δεδομένοι.

δ'. Τῇ θέσει δεδόσθαι λέγονται[2], σημεῖά τε, καὶ γραμμαὶ, καὶ γωνίαι, ἃ τὸν αὐτὸν ἀεὶ τόπον ἐπέχει[3].

DEFINITIONES.

1. Data magnitudine dicuntur, et spatia, et lineæ, et anguli, quibus possumus æqualia invenire.

2. Ratio dari dicitur, cui possumus eamdem invenire.

3. Rectilineæ figuræ specie dari dicuntur, quarum et anguli dati sunt ad unum, et rationes laterum inter se datæ.

4. Positione dari dicuntur, et puncta, et lineæ, et anguli, quæ eumdem semper situm obtinent.

LES DONNÉES

D'EUCLIDE.

1. Des espaces, des lignes, et des angles, auxquels nous pouvons trouver des grandeurs égales, sont dits donnés de grandeur.

2. Une raison est dite donnée, quand nous pouvons lui en trouver une qui soit la même.

3. Des figures rectilignes, dont chacun des angles est donné, et dont les raisons de leurs côtés entre eux sont données, sont dites données d'espèce.

4. Des points, des lignes, et des angles qui conservent toujours la même situation, sont dits donnés de position.

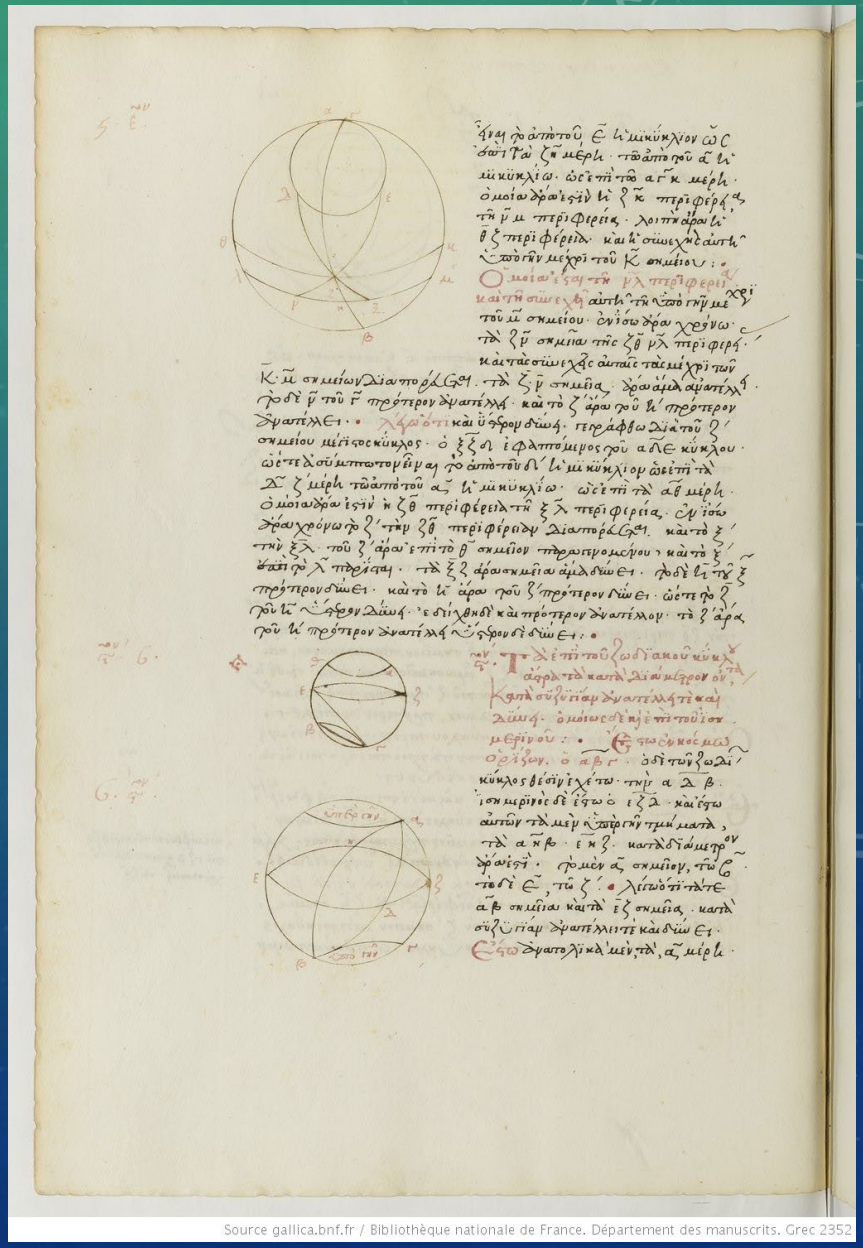
DEDOMENA

In the end, Dedomena is a collection of geometric axioms, 94 in numbers and following a strict pattern:

Given A, than B is also given.

A kind of description or deduction of «known unknowns».

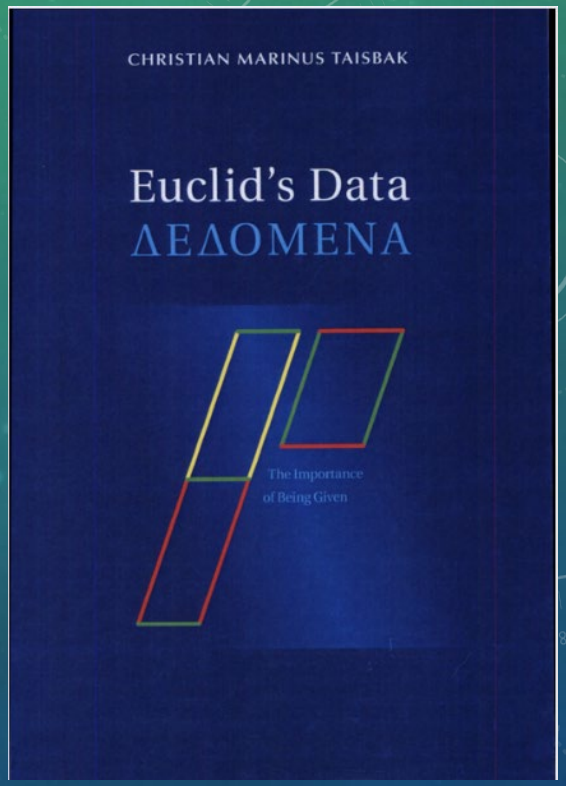
https://gallica.bnf.fr/ark:/12148/btv1b52508678v/f224.item



CHRISTIAN MARINUS TAISBAK

Euclid’s Dedomena. The Importance of Being Given. 2003. p. 14:

When I started to translate the Data, I found it very longwinded that a certain phrase kept popping up time and again, several times in every proposition: if this item is given, that item is also given. I decided to cancel all those alsos… But then I discovered that I was leaving out an essential feature of the “Data”: the Givens hang together in chains, the purpose of any proposition being to produce more links to them.

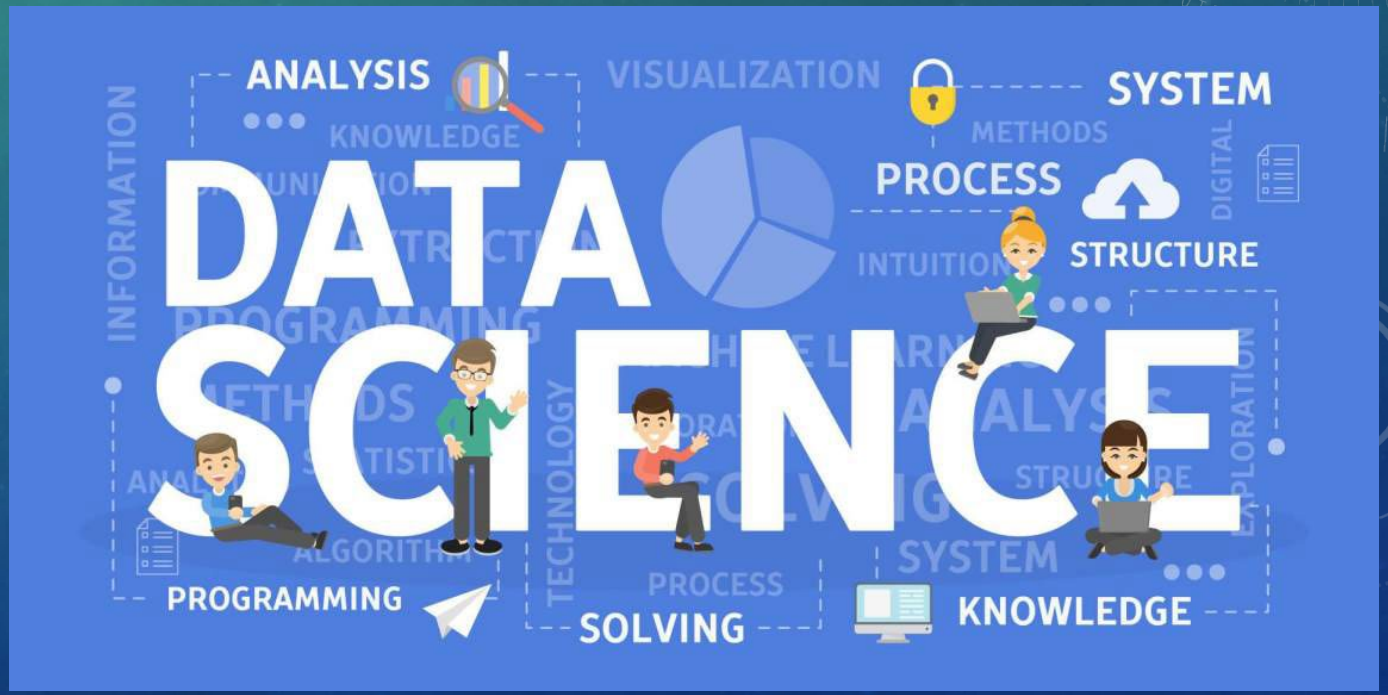


DATA SCIENCE

Today, Data Scientists try to do the same thing

(maybe on a more complex level!)

Some would say, that they try to find the «unknown unknowns».

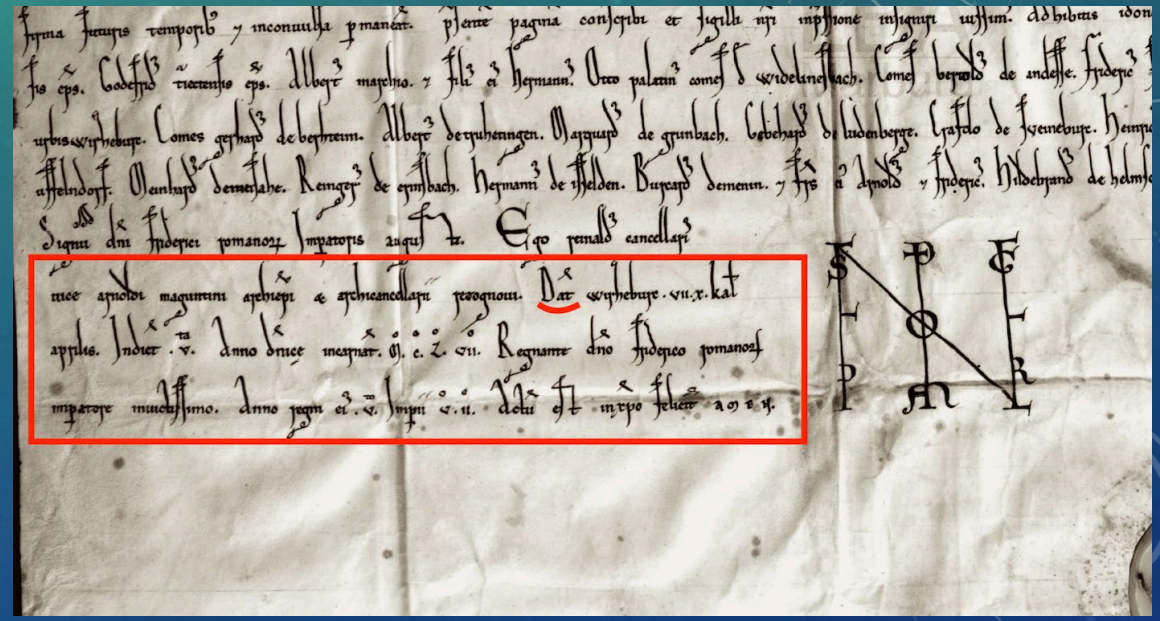


https://thedatascientist.com/data-science-considered-own-discipline/

MIDDLE AGES

During the middle ages, data functions as a time stamp in documents and acts:

Pattern:
This document was created …



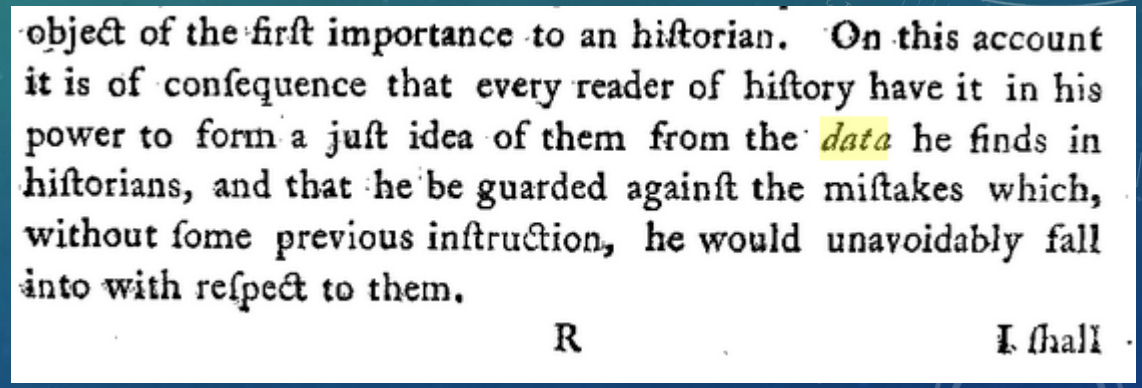
http://lba.hist.uni-marburg.de/lba-cgi/kleioc/0010KlLBA/exec/showrecord/zugangsnummer/%2210234%22

17TH CENTURY

The word data reappears in the middle of the 17th century, interestingly in literature from Mathematics or Divinity.

Data is understood as «anything that is given, either as mathematical assertion or by the word of God».

Soon the term is used in other disciplines.

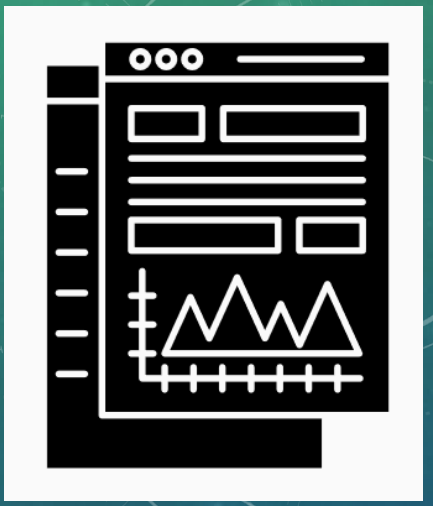


object of the firſt importance to an hiſtorian. On this account it is of conſequence that every reader of hiſtory have it in his power to form a juſt idea of them from the *data* he finds in hiſtorians, and that he be guarded againſt the miſtakes which, without ſome previous inſtruction, he would unavoidably fall into with reſpect to them.

R I ſhall

Joseph Priestley: Lectures on History, and General Policy, 1788, S. 121

18TH CENTURY



At the end of the 18th century, the use of data becomes limited and is only found in relation with experiments, experience or the collection of facts.

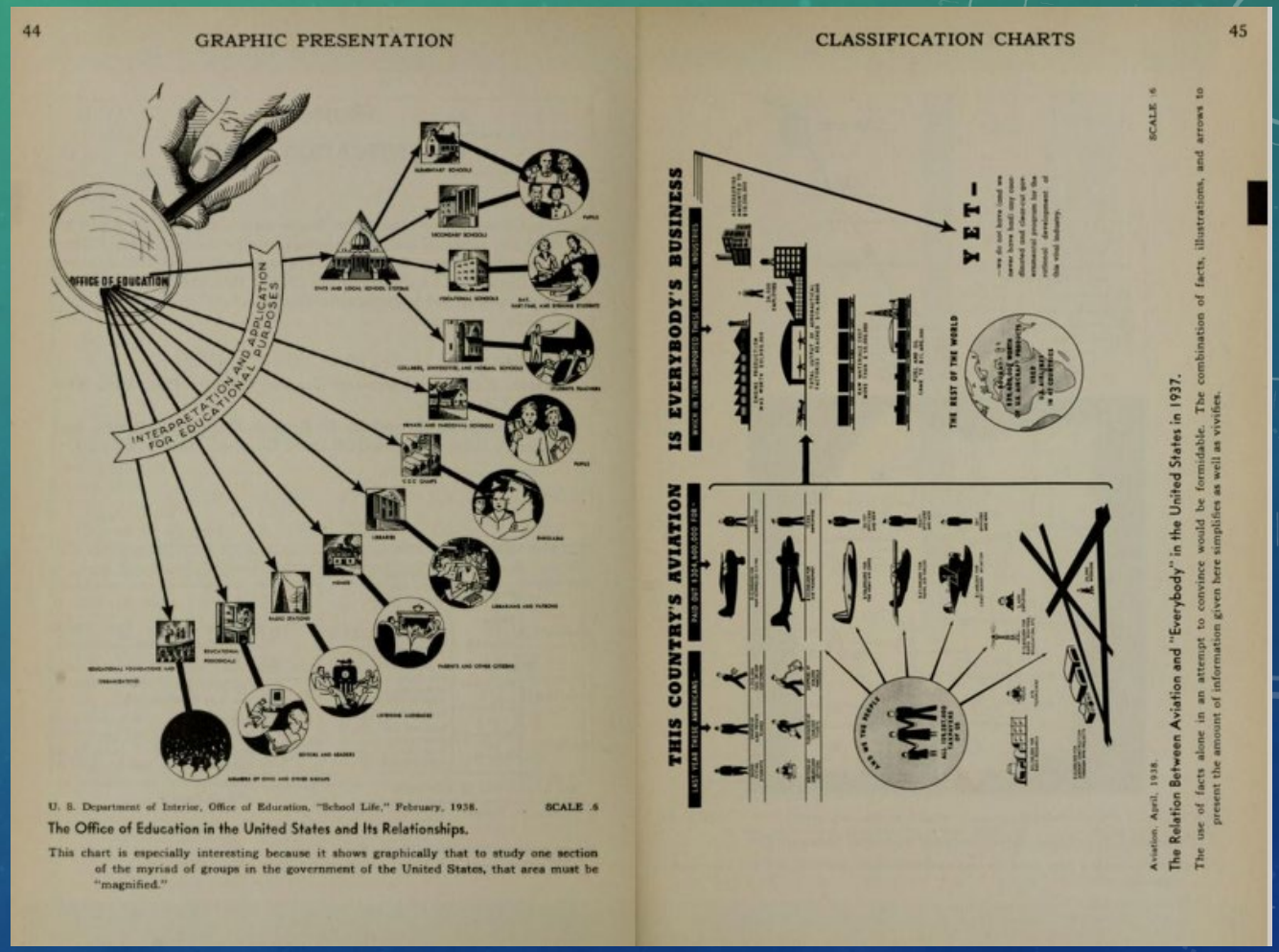
«It had become usual to think of data as the result of an investigation rather than its premise."

Daniel Rosenberg, "Data before the Fact", in: Gitelman, Lisa (ed.): "Raw Data" is an Oxymoron. Cambridge/Mass.: MIT Press, 2013, p. 33.

MODERN TIMES

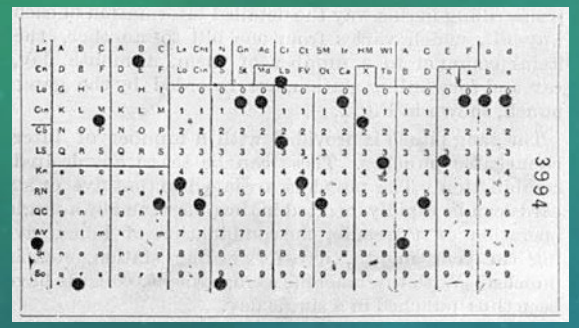
Data in modern times is firstly to be seen as related to descriptive statistics.

And its respective visualization.

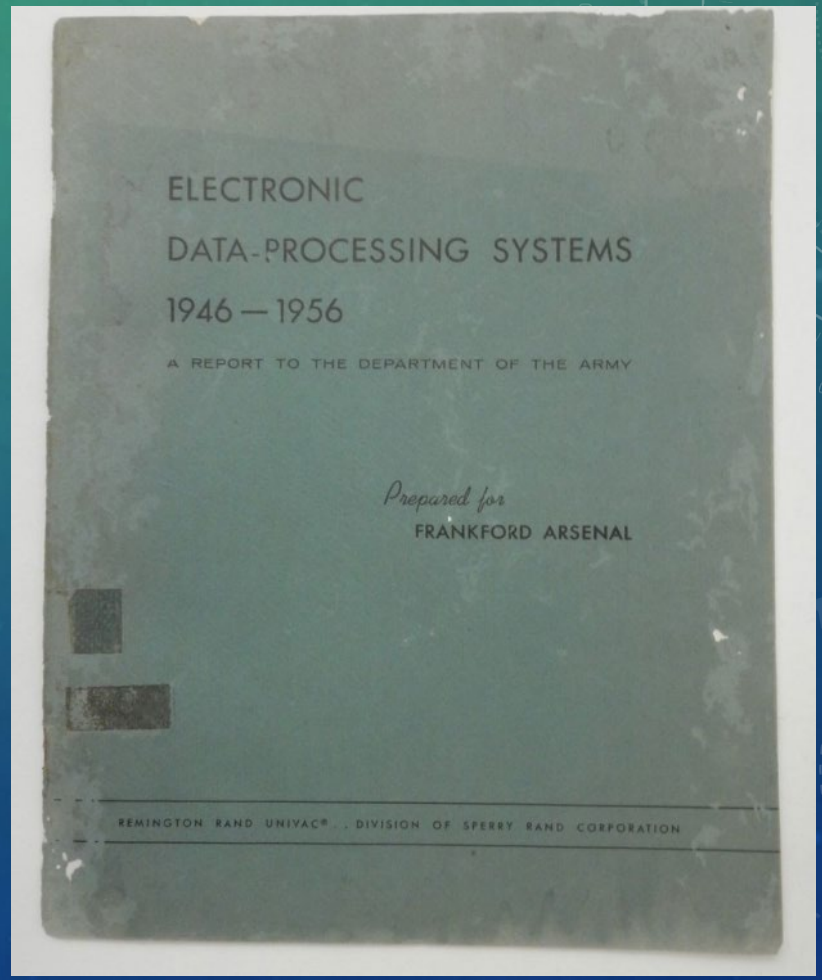


Willard Cope Brinton: Graphic Presentation. 1939.

MODERN TIMES (C'TUED)



Secondly, with the invention of computers, data gains even more importance in the context of what is called Electronic DATA Processing.



Electronic Data-Processing Systems 1946-1956 : A report to the Department of the Army Prepared for Frankford Arsenal. no author.

DATA SCIENCE

The current state of statistical work can be described by a **Statistical Trilogy**:

1. Data Collection (experimental design, sample surveys)

2. Data Modeling and Analysis

3. Problem Understanding/Solving, Decision Making

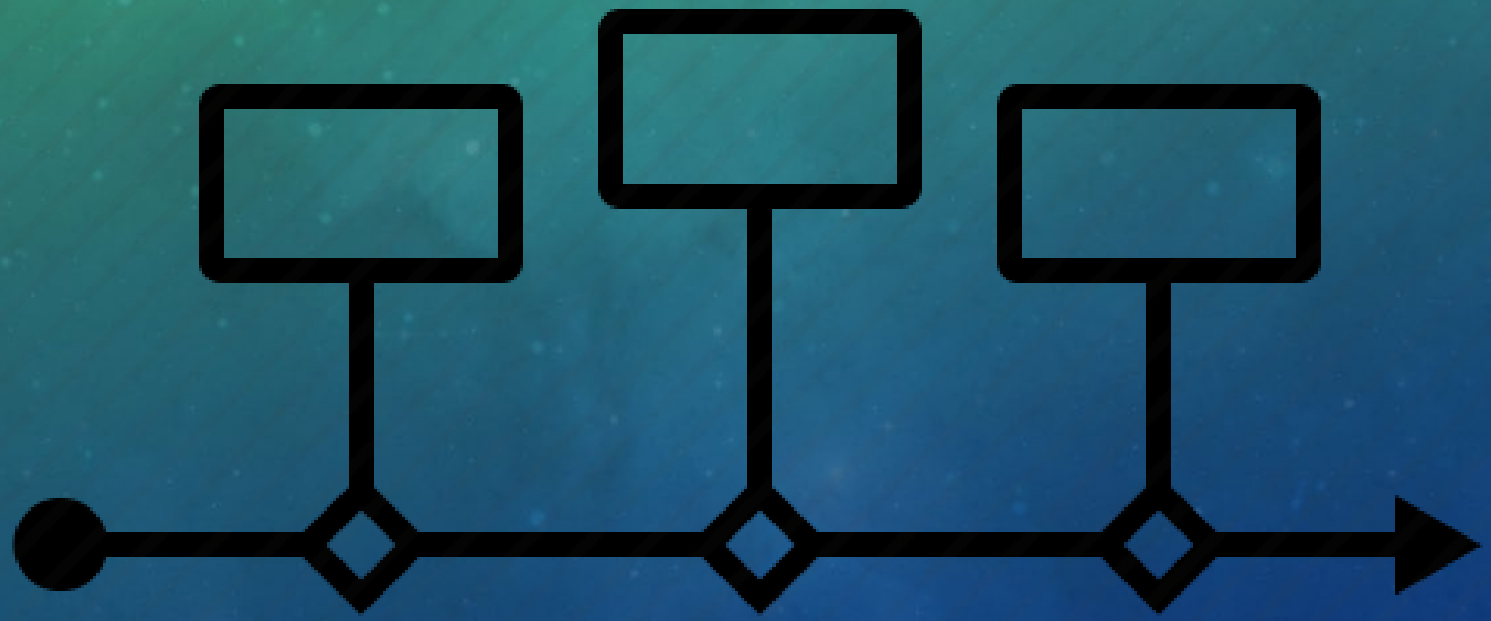
Wu, C. F. J. (1997). "Statistics = Data Science?" (PDF).

DATA OVER TIME

DESCRIPTIVE

PRESCRIPTIVE

PREDICTIVE



PRESENCE: DATA EVERYWHERE

Data Word Cloud, Business Concept Royalty…
123rf.com

big data word cloud – Magpi
home.magpi.com

Word Cloud of Big Data - BusiTelCe
busitelce.com

Visualizing Text Analysis Results with Word C…
dundas.com

Word cloud of Master Data Management | Downl…
researchgate.net

Word Cloud Big Data Stock Illustration - D…
stockphoto.com

big-data-word-cloud - Resolve Solution Partners
resolvesp.com

Word cloud data integration. Word clo…
canstockphoto.com

DATA word cloud collage, bu…
colourbox.com

Big Data Word Cloud Collage Techn…
depositphotos.com

Big Data Word Cloud Stock Vecto…
shutterstock.com

WHAT WE SEE SO FAR… IS THAT

The term data has a long history. As long as science.

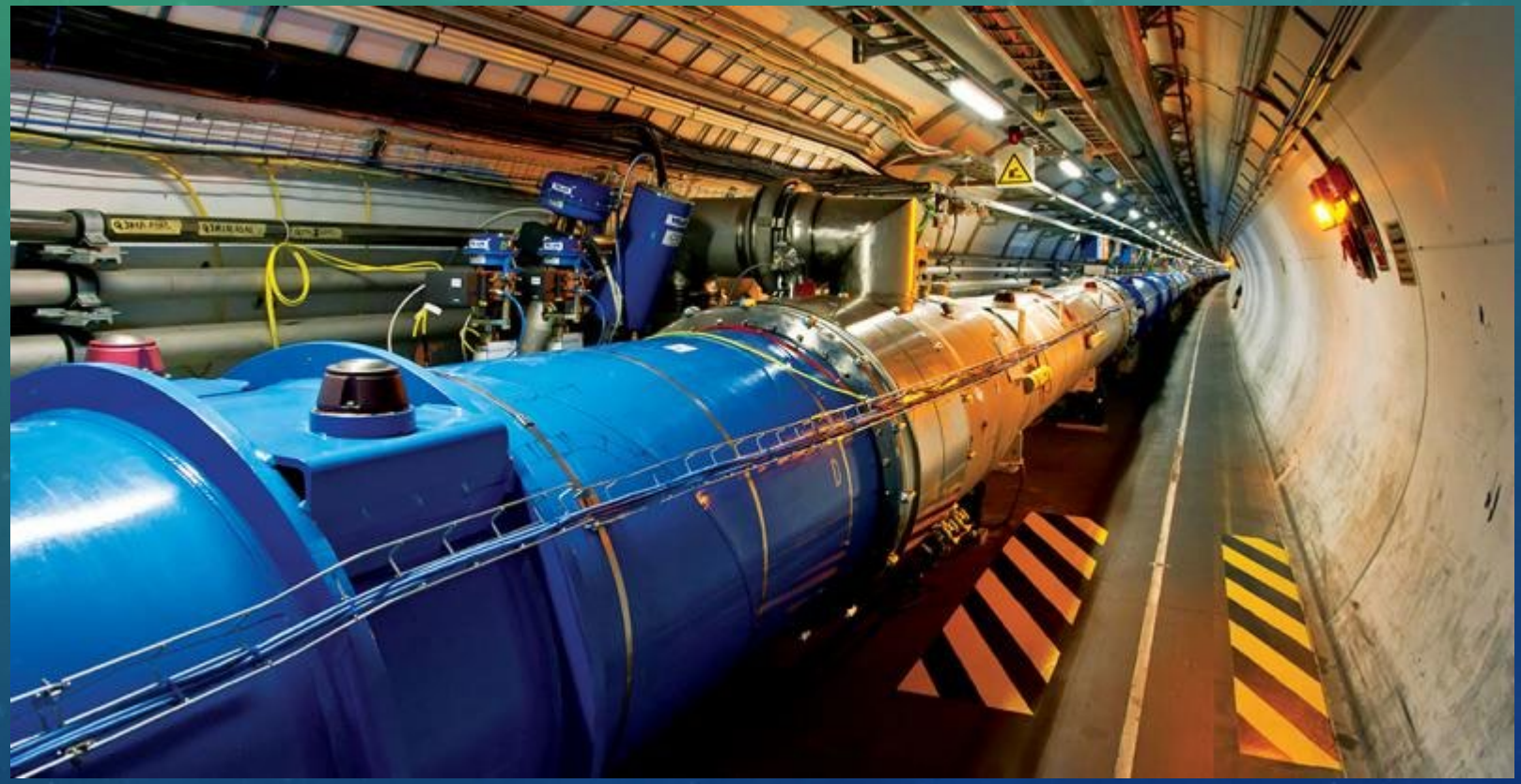
Data do have many facets.

On facet is research data. And its management.

RESEARCH DATA - DEFINITON

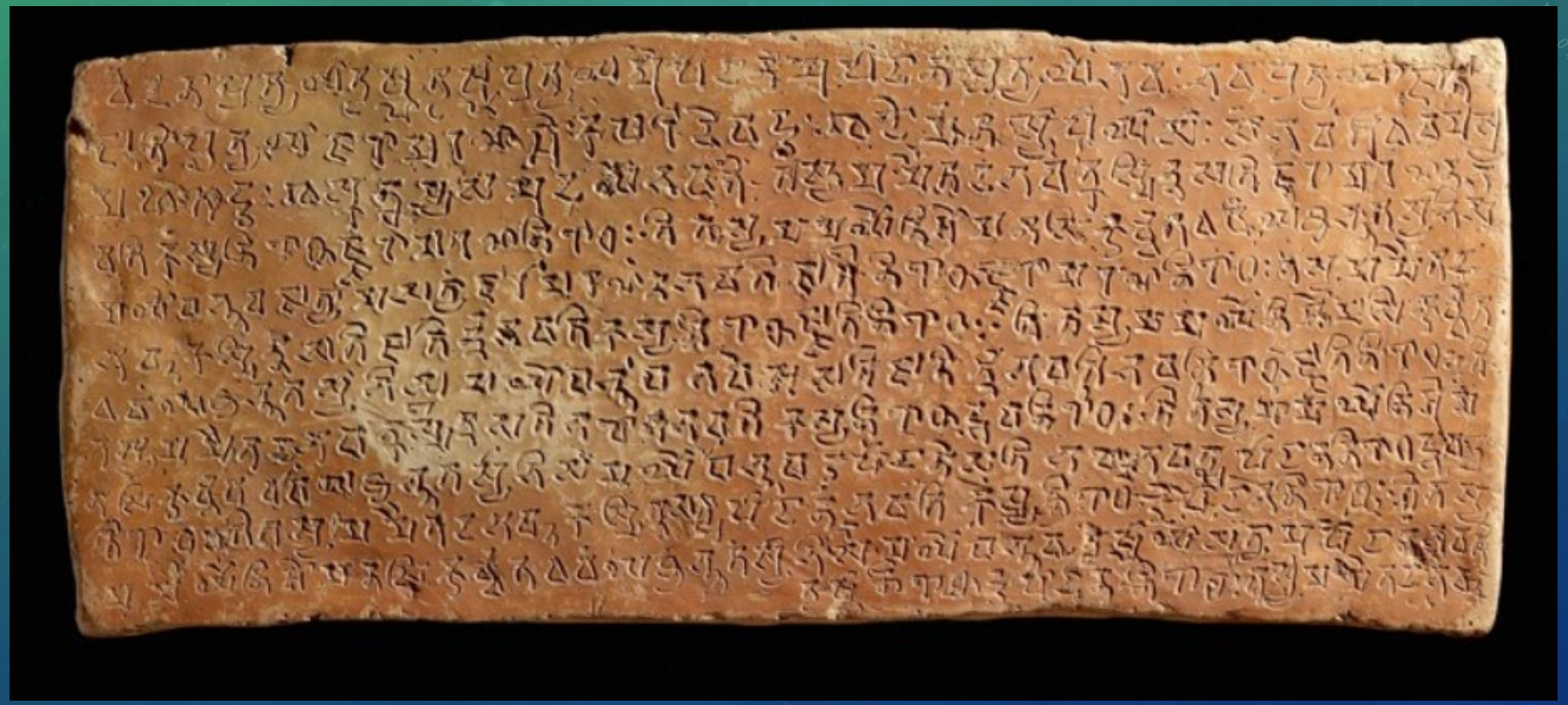
- « any information in binary digital form». Includes digital objects and data bases. (Harvey Ross. Digital Curation. 2010)
- Research data are collections, observations, models, measurements, references, digitizations etc.
- The definition of research data strongly depends on the scientific or disciplinary context of their creation and the scientific perspectives taken or methods used.

RESEARCH DATA EXAMPLE 1



Large Hadron Collider, CERN 1) Raw 2) Reconstructed 3) Reduced 4) Published

RESEARCH DATA EXAMPLE 2



Brick inscribed with the sutra of dependent origination
Gorakhpur, 5-6ième siècle

RESEARCH DATA: EXAMPLE 3
STANFORD MARSHMALLOW PROJECT - 1960

STANFORD MARSHMALLOW PROJECT - REVISION

1988

1990

2006

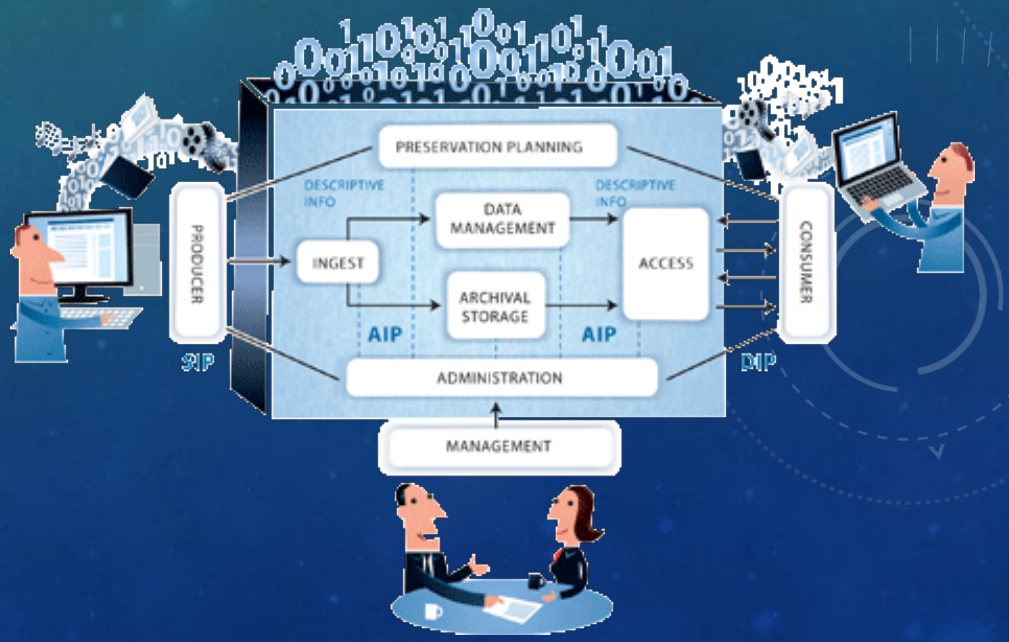
2011 …

-> The data collected in 1960 must have been stored.

But how? On which medium? In which format?

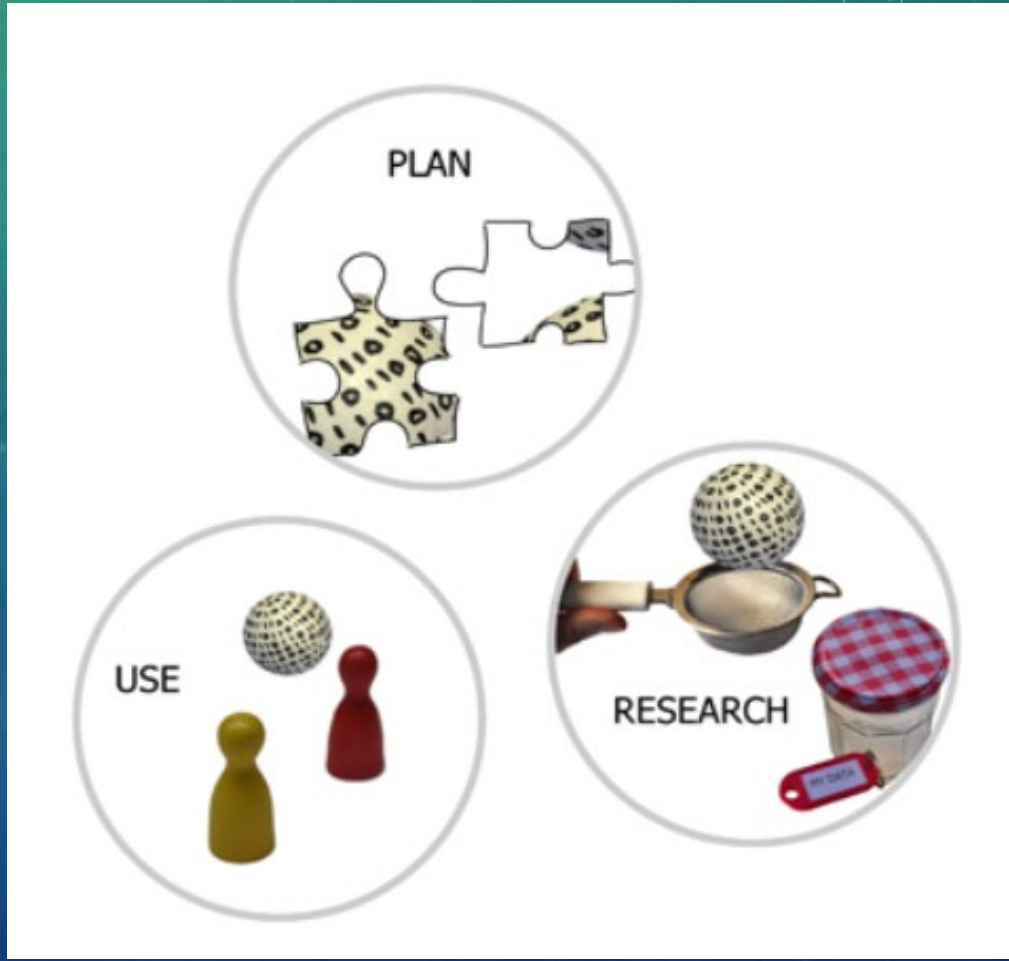
2

Χάρυβδις
The data life cycle



SIMPLE CYCLE

http://datasupport.researchdata.nl



DATA CURATION

“active and on-going management of data through its lifecycle of interest and usefulness to scholarship, science, and education” through activities that “enable data discovery and retrieval, maintain quality, add value, and provide for re-use over time.”

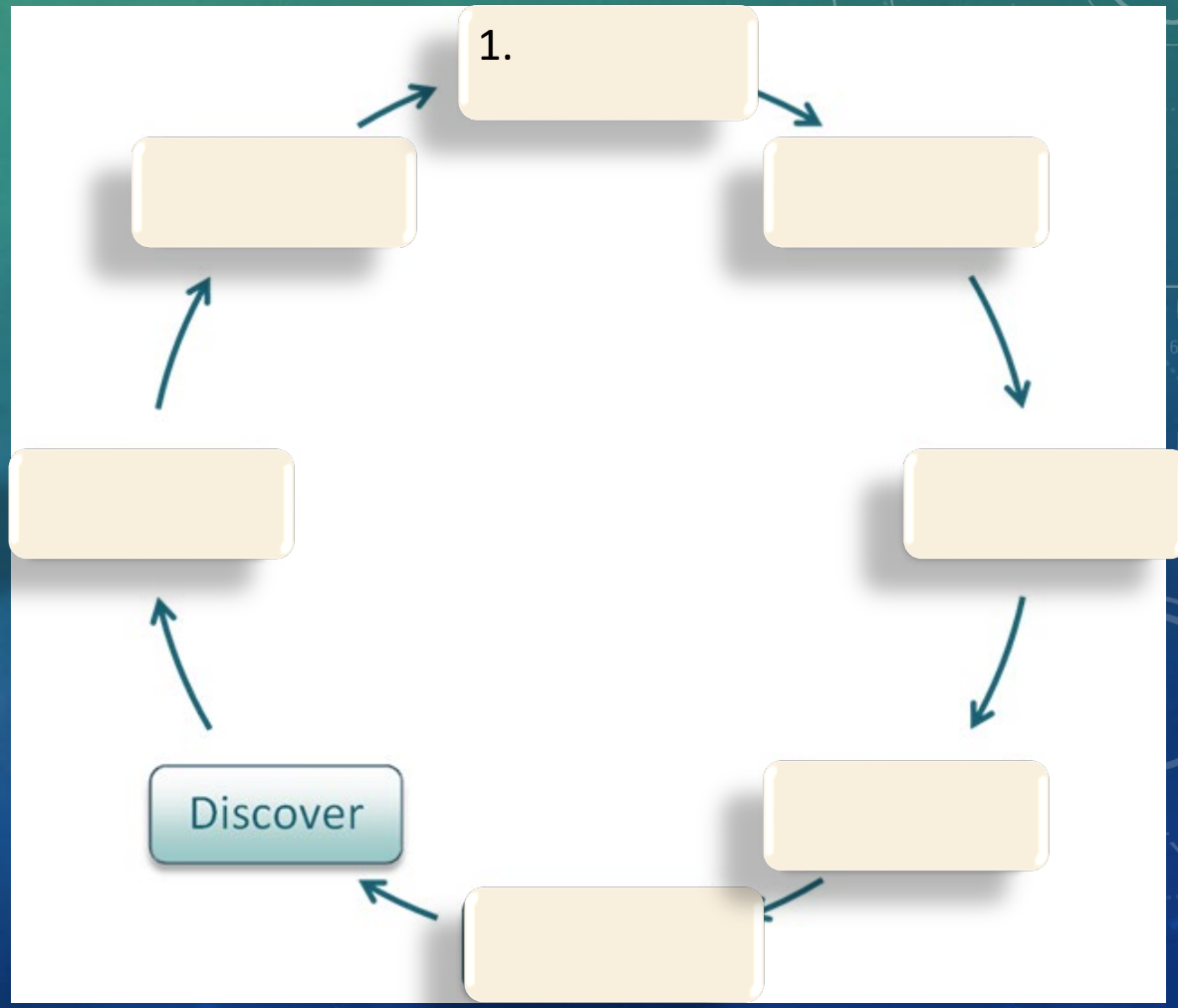
http://hdl.handle.net/2142/3493

EXERCICE LIFE CYCLE

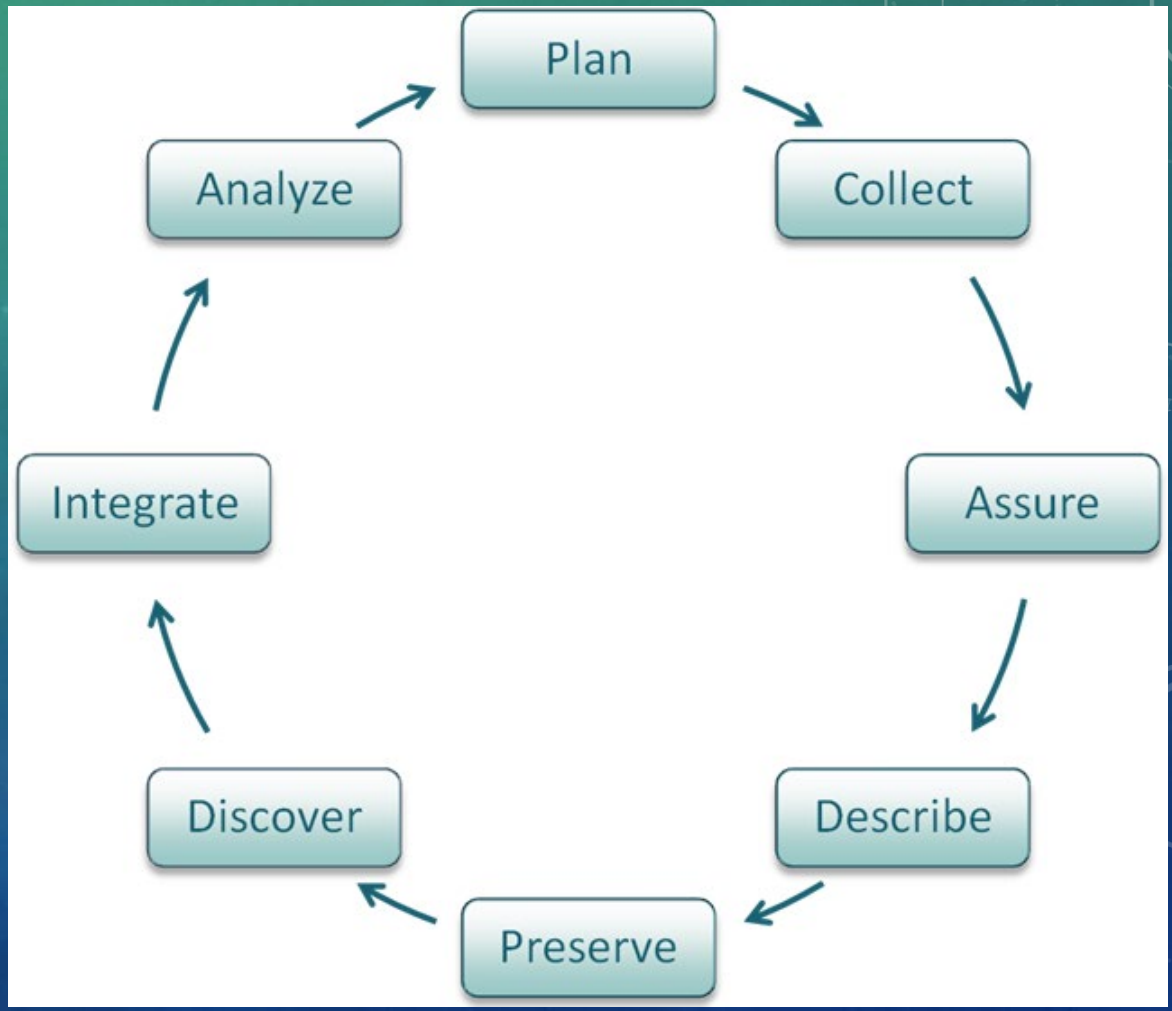
Try to put the actions of the data curation life cycle in the right order!

RESEARCH DATA LIFE CYCLE

- Describe
- Collect
- Integrate
- Analyze
- Plan
- Preserve
- Assure

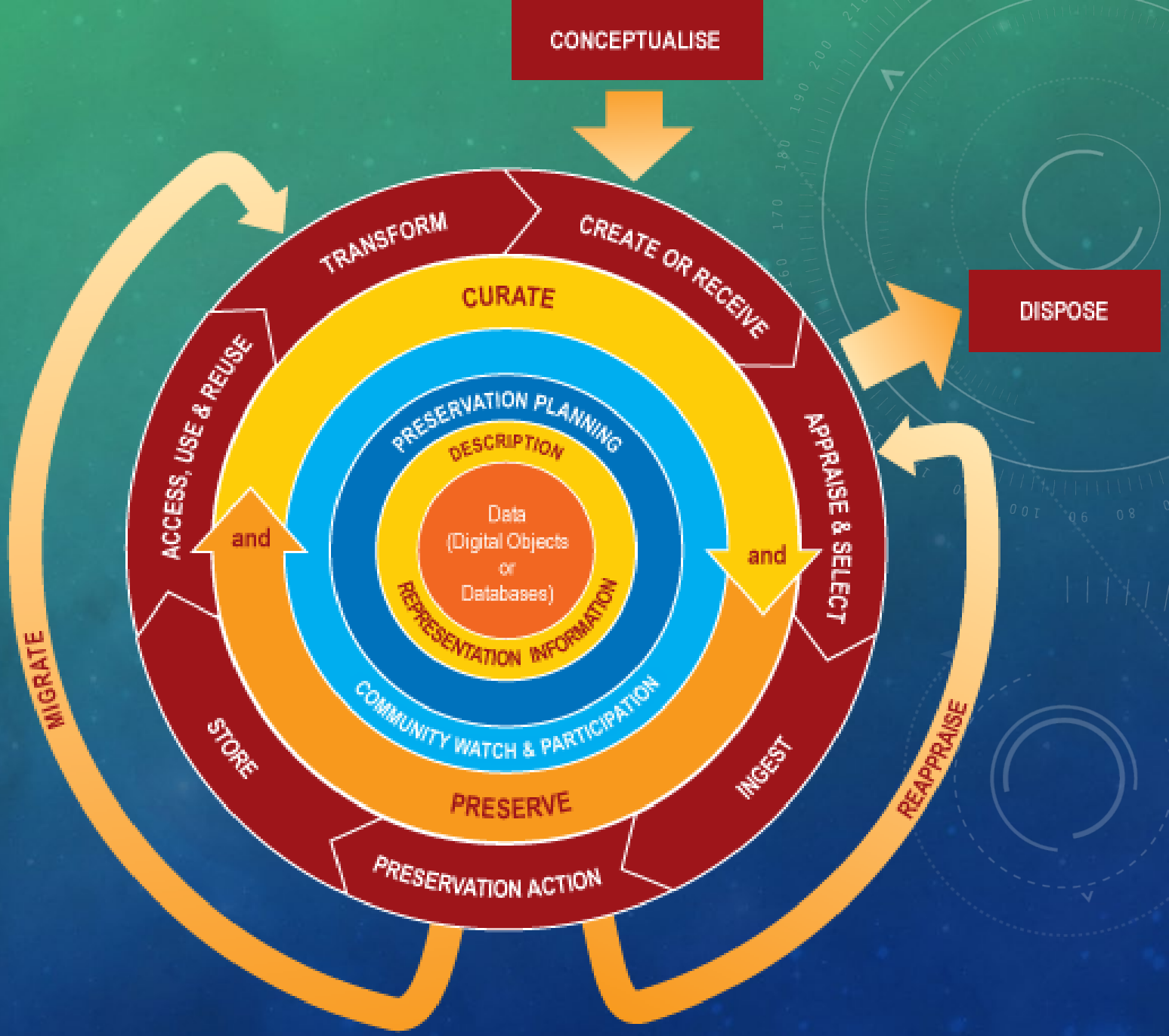


RESEARCH DATA LIFE CYCLE



Reference model Data ONE Project
https://www.dataone.org/best-practices

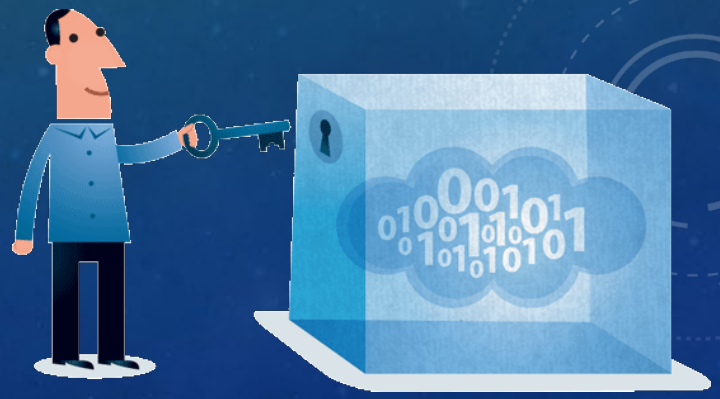
THE DCC CURATION LIFECYCLE MODEL



http://www.dcc.ac.uk/resources/curation-lifecycle-model

3

Σκύλλα
The data continuum model

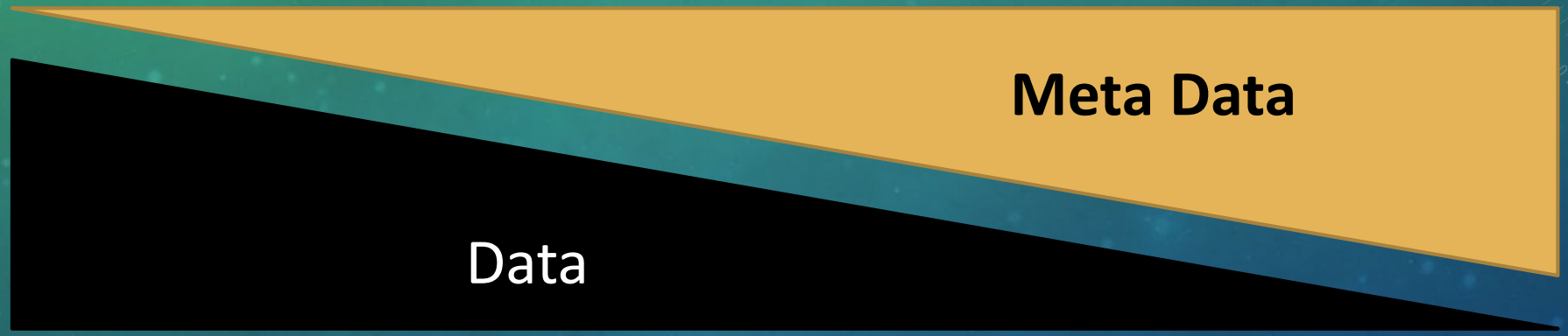


DATA CONTINUA

Object:	Less Metadata	⟷	More Metadata
	More Items	⟷	Fewer Items
	Larger Objects	⟷	Smaller Objects
	Objects continually updated	⟷	Objects static/derived snapshots
Management:	Researcher Manages	⟷	Organisation Manages
	Less Preservation	⟷	More Preservation
Access:	Mostly Closed Access	⟷	Mostly Open Access
	Less Exposure	⟷	More Exposure

http://www.dlib.org/dlib/september07/treloar/treloar-table1.png

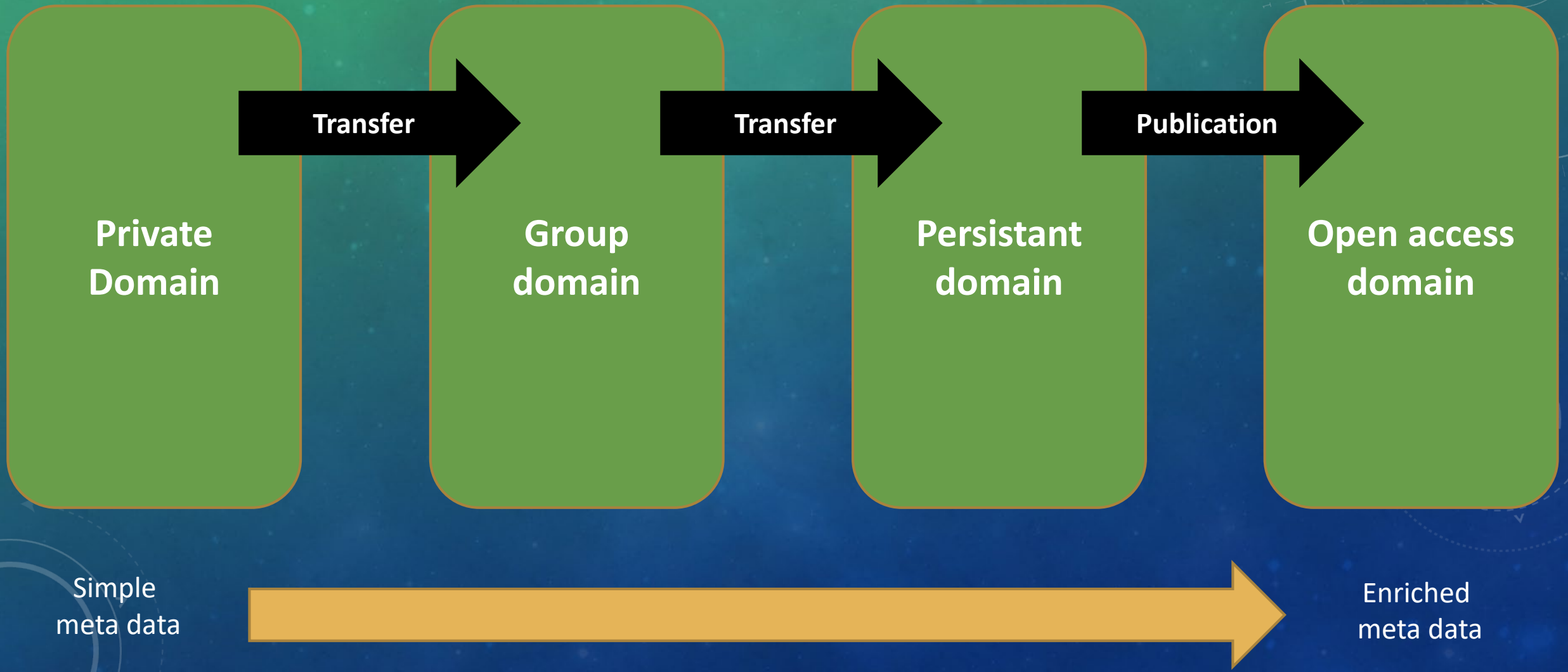
DATA AND METADATA



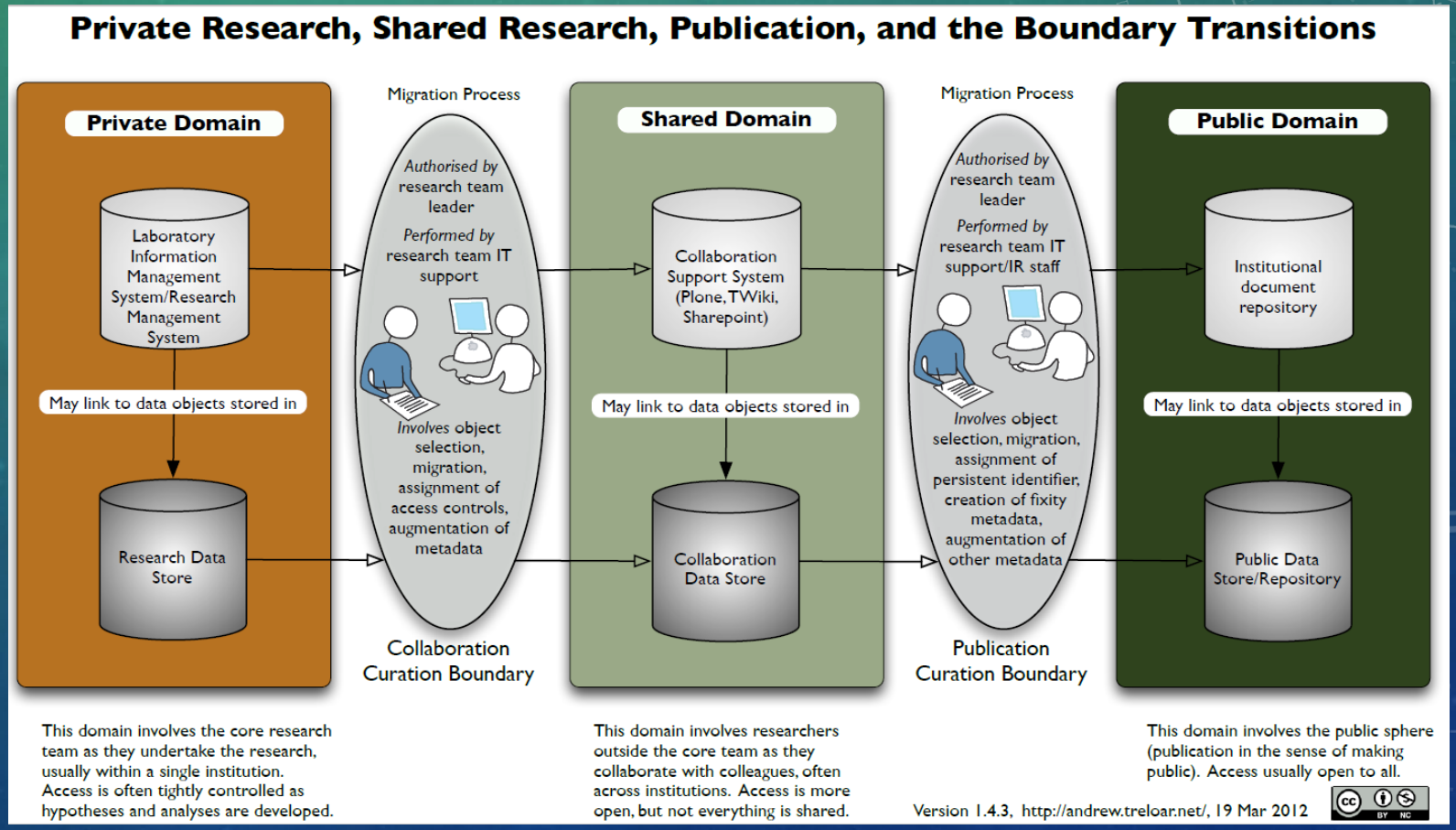
Simple Metadata

Enriched Metadata

DATA CONTINUUM MODEL

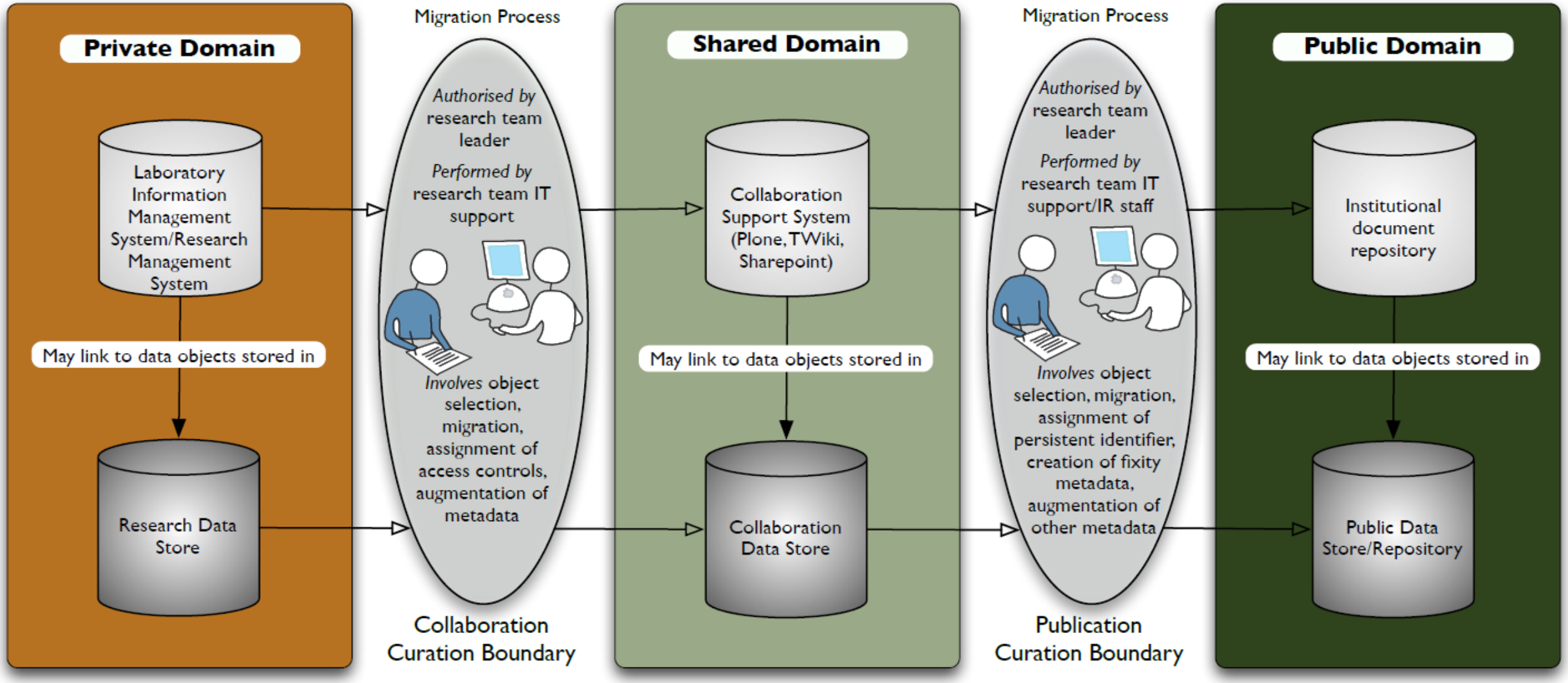


DATA CONTINUUM MODEL



http://andrew.treloar.net/research/diagrams/

Private Research, Shared Research, Publication, and the Boundary Transitions



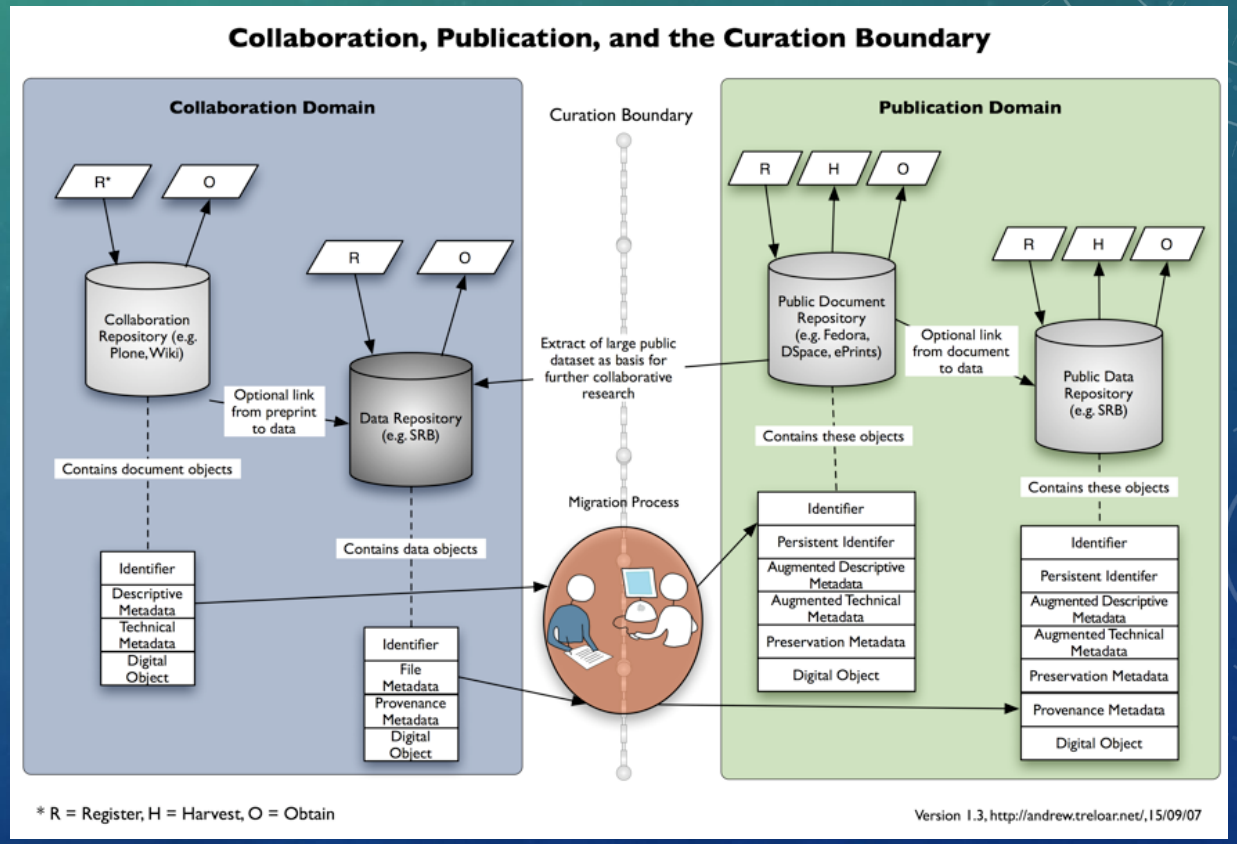
This domain involves the core research team as they undertake the research, usually within a single institution. Access is often tightly controlled as hypotheses and analyses are developed.

This domain involves researchers outside the core team as they collaborate with colleagues, often across institutions. Access is more open, but not everything is shared.

This domain involves the public sphere (publication in the sense of making public). Access usually open to all.

Version 1.4.3, http://andrew.treloar.net/, 19 Mar 2012

COLLABORATION, PUBLICATION AND THE CURATION BOUNDARY

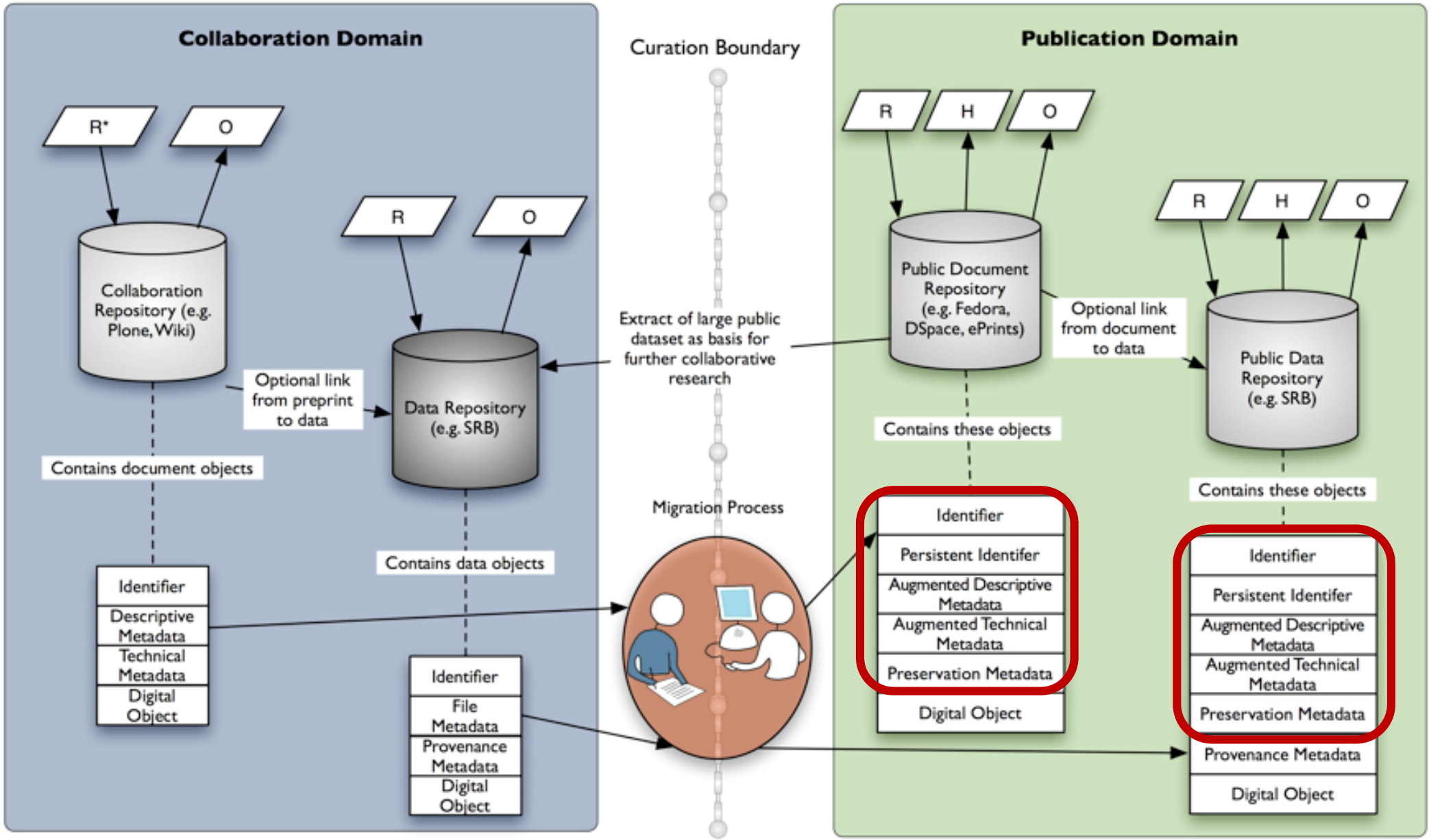


COLLABORATION, PUBLICATION AND THE CURATION BOUNDARY

“These domains are separated by “Curation boundaries”, which are virtual decision points at which the creators of data decide what they will share, with whom, with what metadata and under what conditions.”

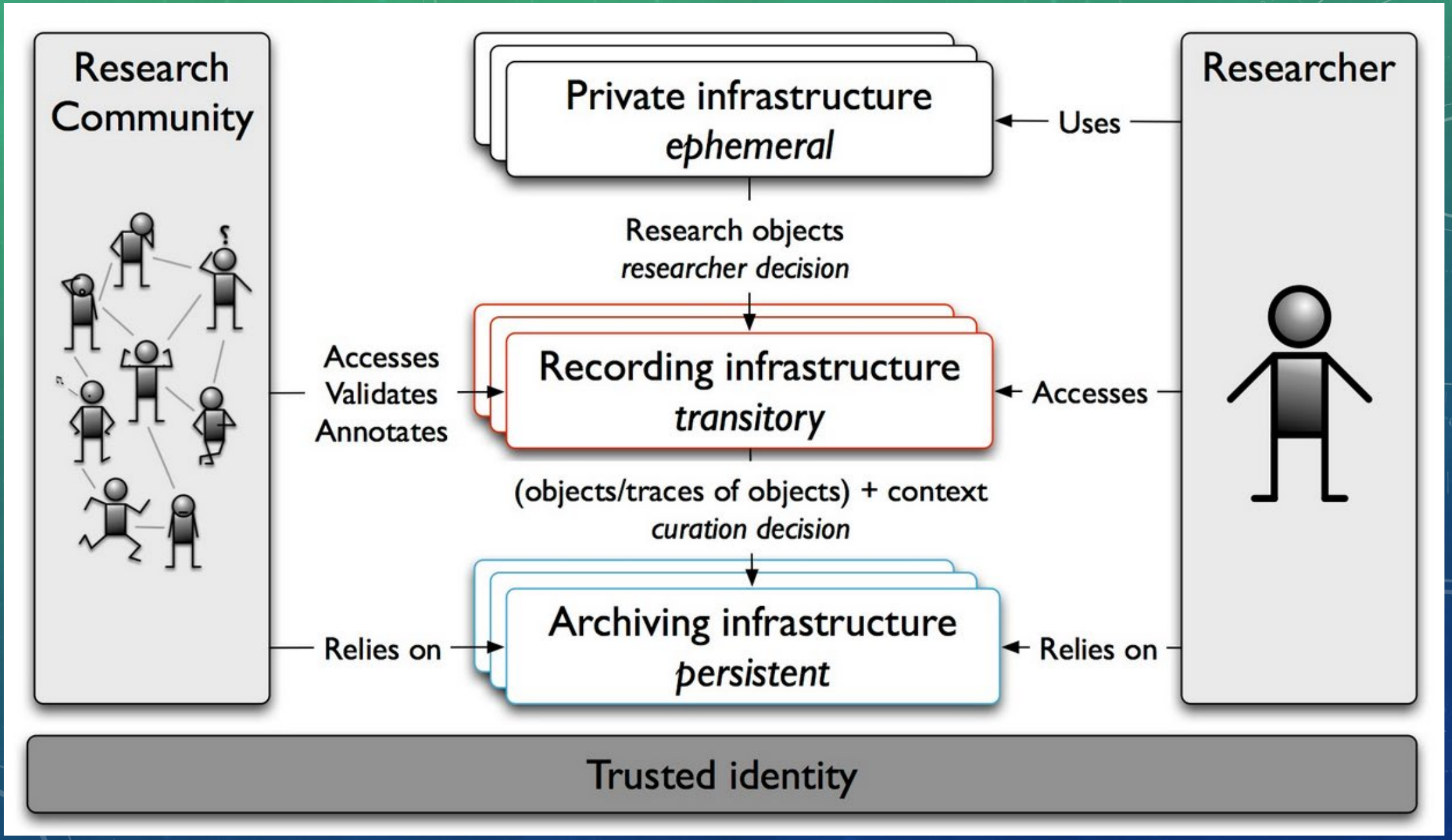
https://figshare.com/articles/figshare_and_Monash_University_combining_cloud_management_and_discoverability_with_institutional_storage/1224755

Collaboration, Publication, and the Curation Boundary



* R = Register, H = Harvest, O = Obtain

Version 1.3, http://andrew.treloar.net/, 15/09/07

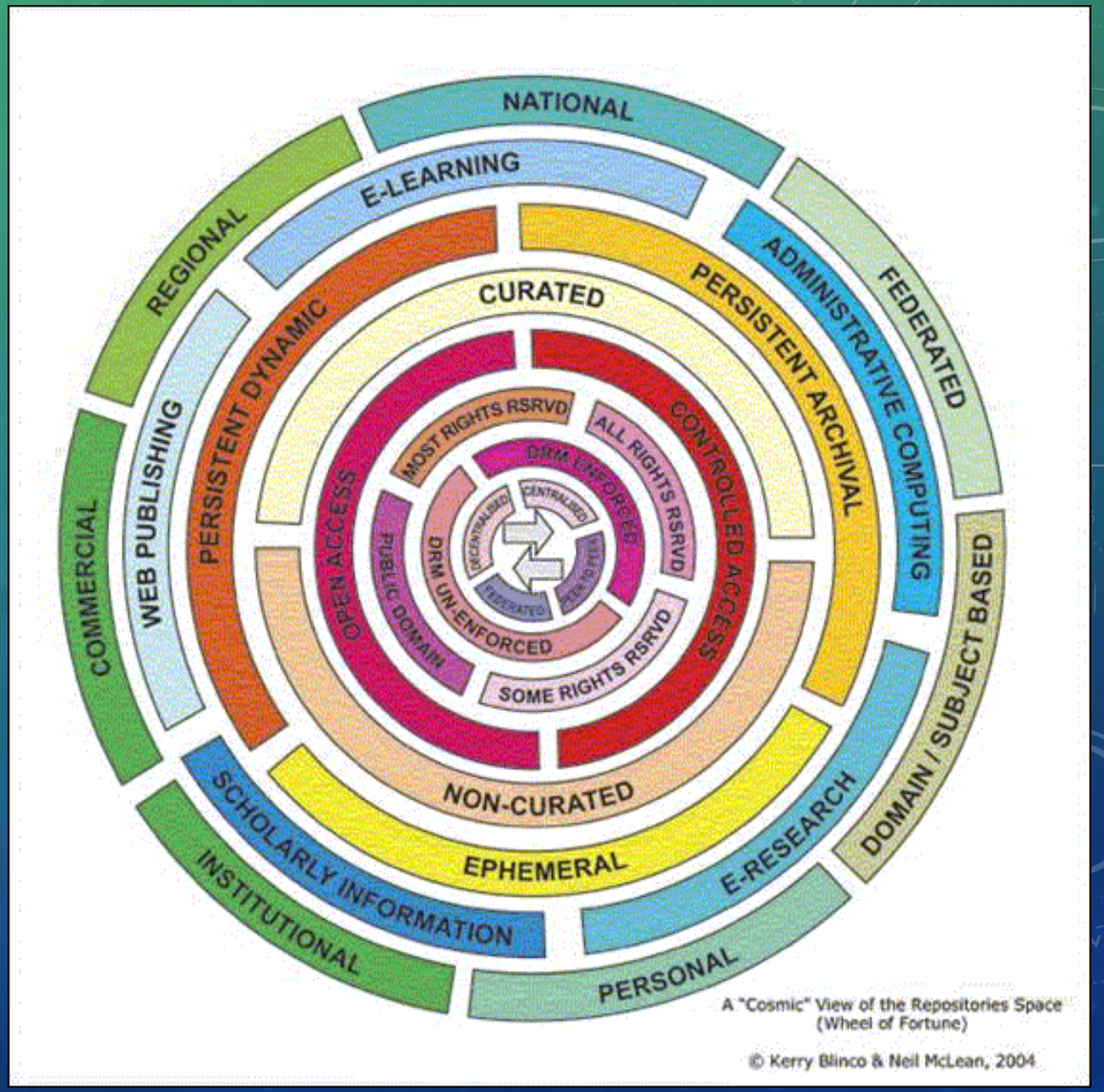


http://andrew.treloar.net/research/diagrams/recording-to-archiving-architecture.jpg

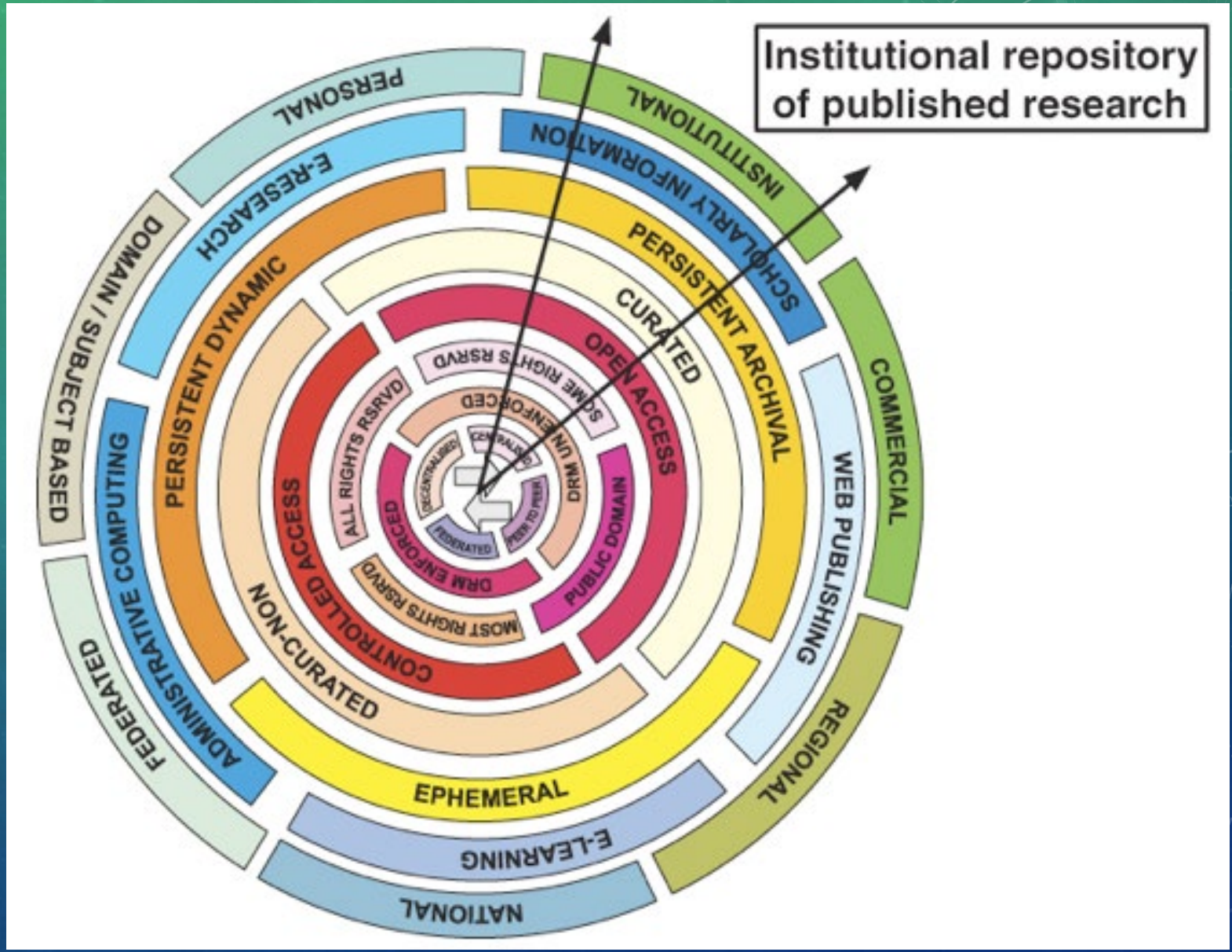
EXERCICE REPOSITORIES

Take the Wheel of Fortune in your hands!

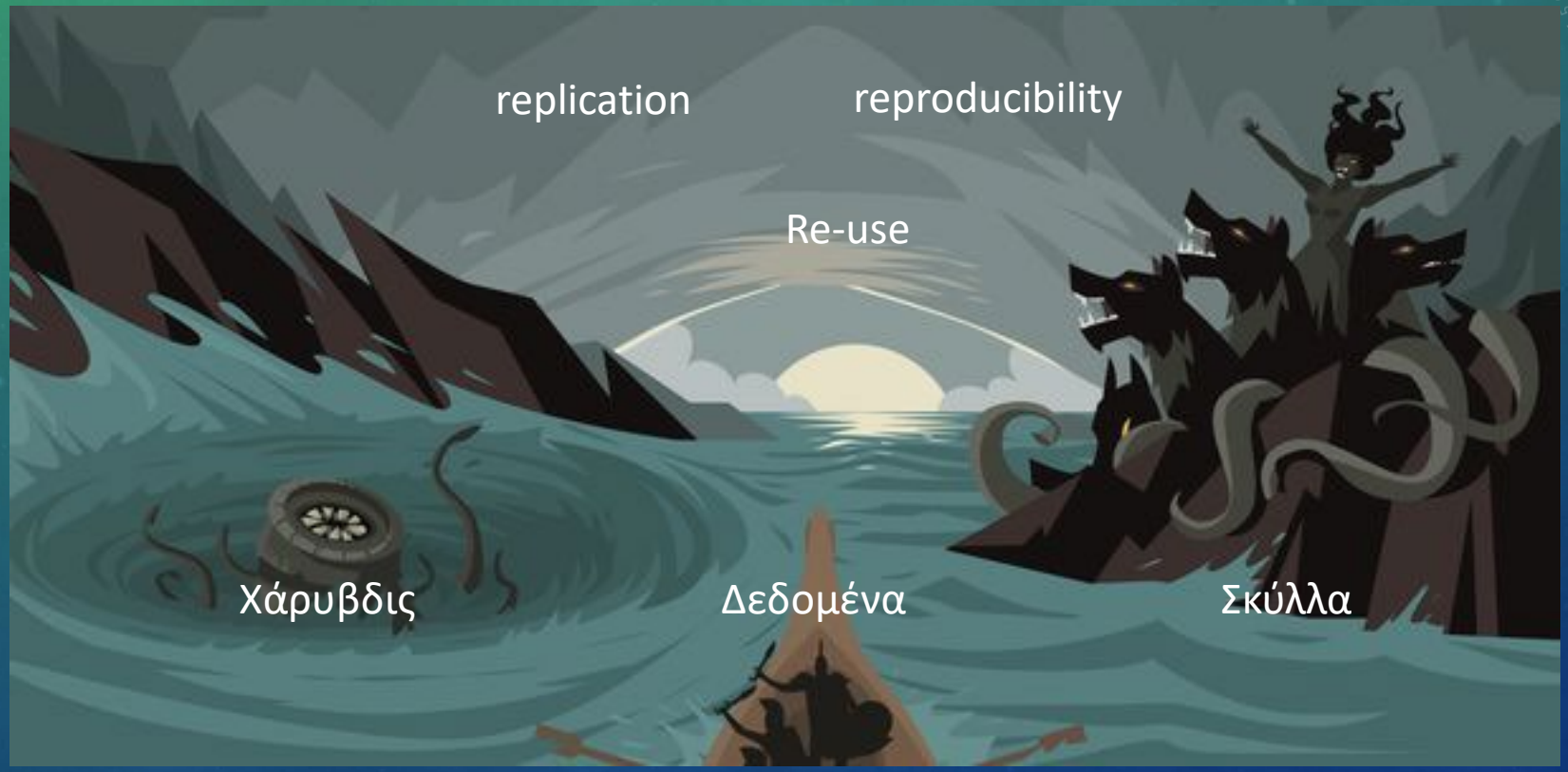
What is the position of each wheel to create an Institutional Open Access Repository for research data?


NATIONAL
E-LEARNING
REGIONAL
FEDERATED
ADMINISTRATIVE COMPUTING
COMMERCIAL
WEB PUBLISHING
PERSISTENT DYNAMIC
CURATED
PERSISTENT ARCHIVAL
MOST RIGHTS RSRVD
ALL RIGHTS RSRVD
DRM ENFORCED
CONTROLLED ACCESS
OPEN ACCESS
PUBLIC DOMAIN
DRM UN-ENFORCED
SOME RIGHTS RSRVD
NON-CURATED
EPHEMERAL
SCHOLARLY INFORMATION
E-RESEARCH
INSTITUTIONAL
PERSONAL
DOMAIN / SUBJECT BASED
A "Cosmic" View of the Repositories Space
(Wheel of Fortune)
© Kerry Blinco & Neil McLean, 2004

SOLUTION


Institutional repository of published research
PERSONAL
INSTITUTIONAL
E-RESEARCH
SCHOLARLY INFORMATION
DOMAIN / SUBJECT BASED
PERSISTENT DYNAMIC
PERSISTENT ARCHIVAL
CURATED
COMMERCIAL
OPEN ACCESS
SOME RIGHTS RSRVD
ALL RIGHTS RSRVD
DRM UNENFORCED
WEB PUBLISHING
CONTROLLED ACCESS
PUBLIC DOMAIN
DRM ENFORCED
MOST RIGHTS RSRVD
NON-CURATED
FEDERATIVE COMPUTING
ADMINISTRATIVE
REGIONAL
EPHEMERAL
E-LEARNING
NATIONAL

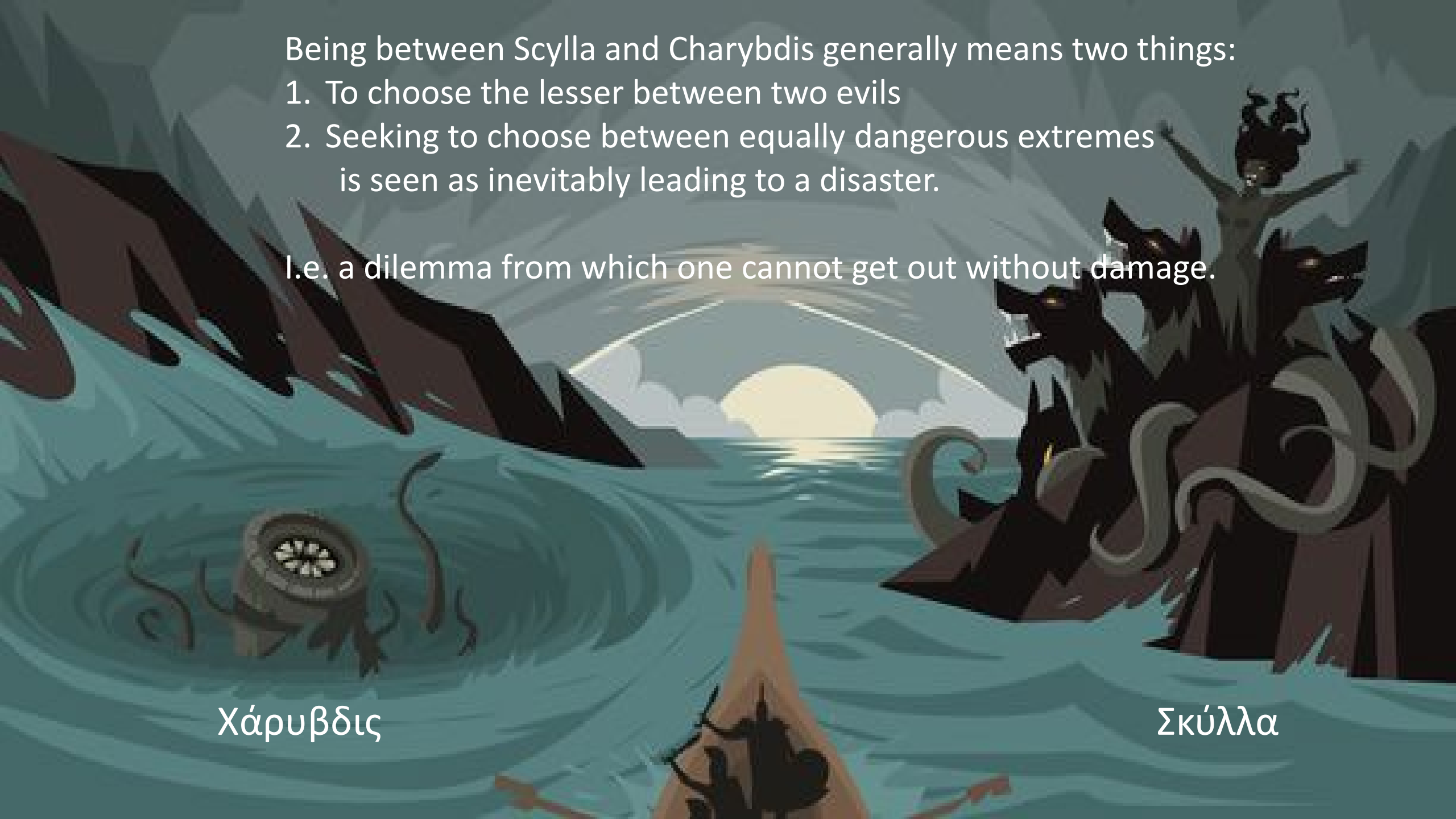
CONCLUSIONS




Σκύλλα
Χάρυβδις
life cycle
continuum
Δεδομένα


Σκύλλα
Χάρυβδις
« as long as possible »
« as soon as possible »
Δεδομένα


Σκύλλα
Χάρυβδις
Long Term Archive
Open Access Repository
Δεδομένα



Being between Scylla and Charybdis generally means two things:

1. To choose the lesser between two evils
2. Seeking to choose between equally dangerous extremes is seen as inevitably leading to a disaster.

I.e. a dilemma from which one cannot get out without damage.

Χάρυβδις

Σκύλλα

What does this mean for us?
If you go for re-use you will either

- Be attracted by a more open access (repository) approach

or

- be opting for a long term archival approach.

Find that out first!
Γνῶθι σεαυτόν

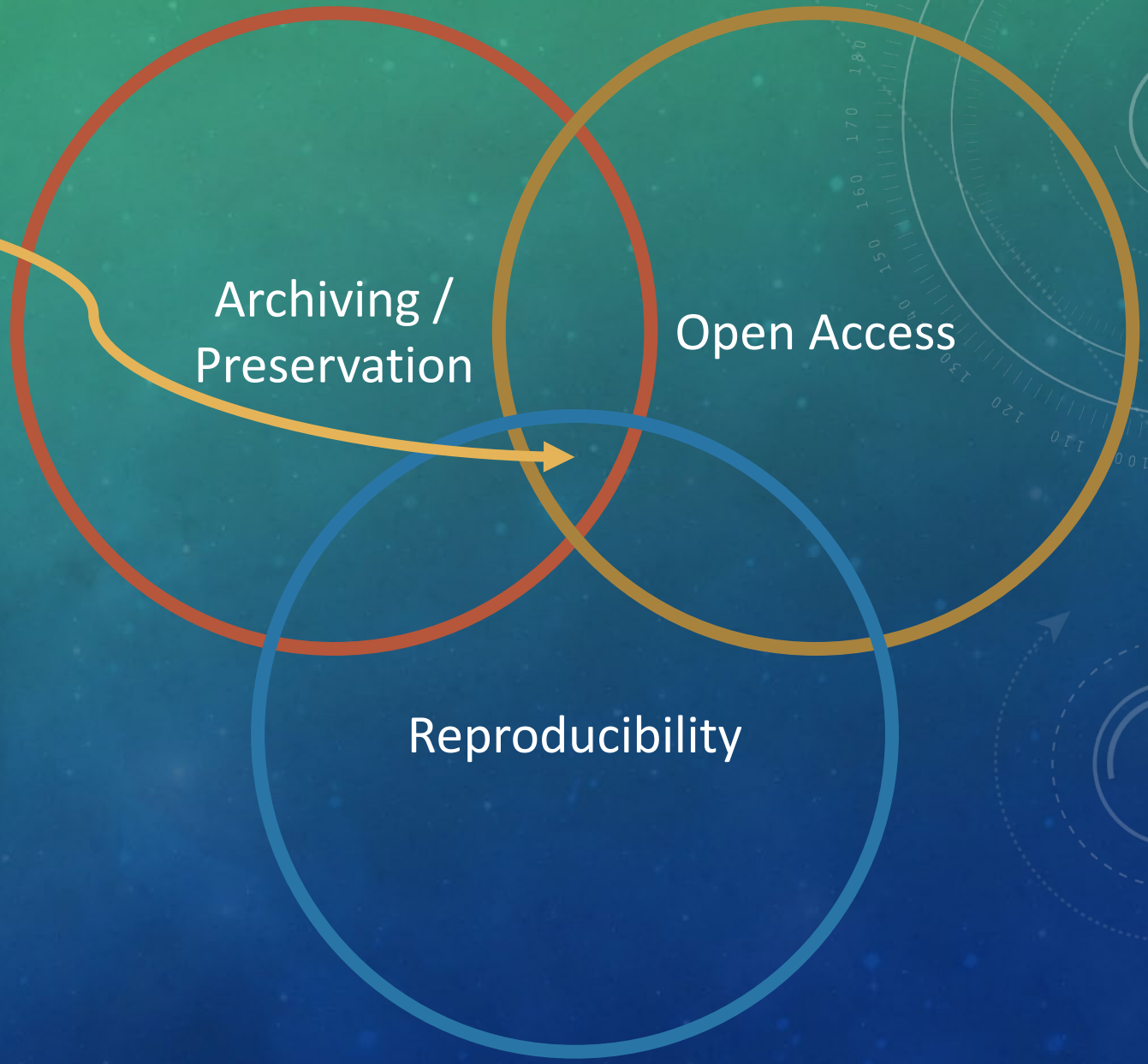
None of the approaches is perfect!
Be aware that each one,
the « as long as possible» approach
as well as
« as soon as possible» approach
has advantages and disadvantages.

The only question might be: Is one of them closer to reproducibility or replication?


replication
reproducibility
re-use
Χάρυβδις
Δεδομένα
Σκύλλα

SWEET SPOT

OF OPEN SCIENCE



LITERATURE 1ST PART

Sandra Rendgen: What do we mean by «data»? Idalab. Blog.

https://idalab.de/blog/data-science/what-do-we-mean-by-data

Daniel Rosenberg, "Data before the Fact", in: Gitelman, Lisa (ed.): "Raw Data" is an Oxymoron. Cambridge/Mass.: MIT Press, 2013, p. 33.

LITERATURE 2ND AND 3RD PART

Harvey Ross. Digital Curation. A How-To-Do-It-Manual. Neal Schuman Publishers, 2010.

Treloar, Andrew, David Groenewegen, and Cathrine Harboe-Ree. "The data curation continuum: Managing data objects in institutional repositories." D-Lib magazine 13.9 (2007): 4.