Open Science and Open Data
data types & standards

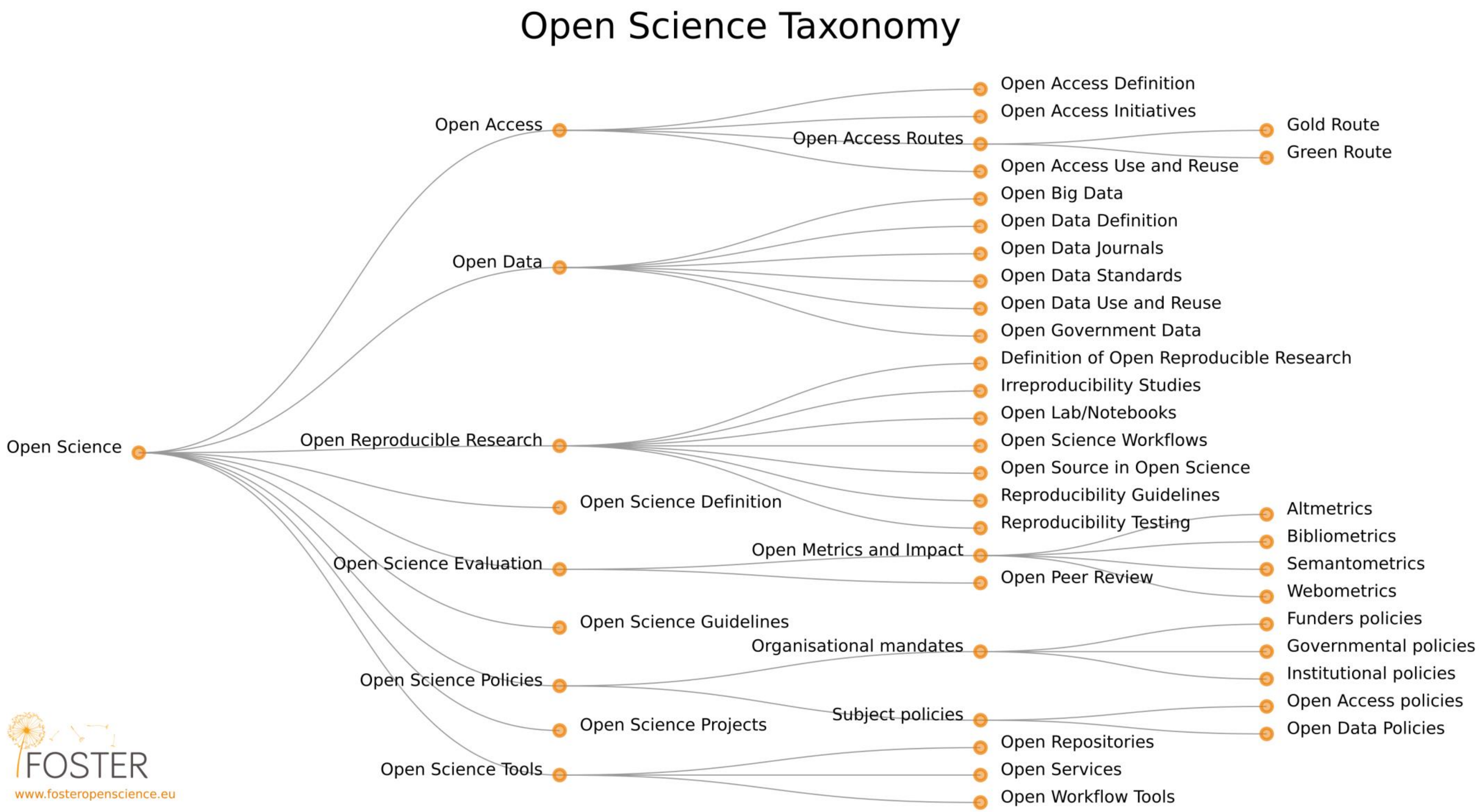
~

Biodiversity data management and open data
ForBIO Course
Tartu - 2016

Hanna Koivula

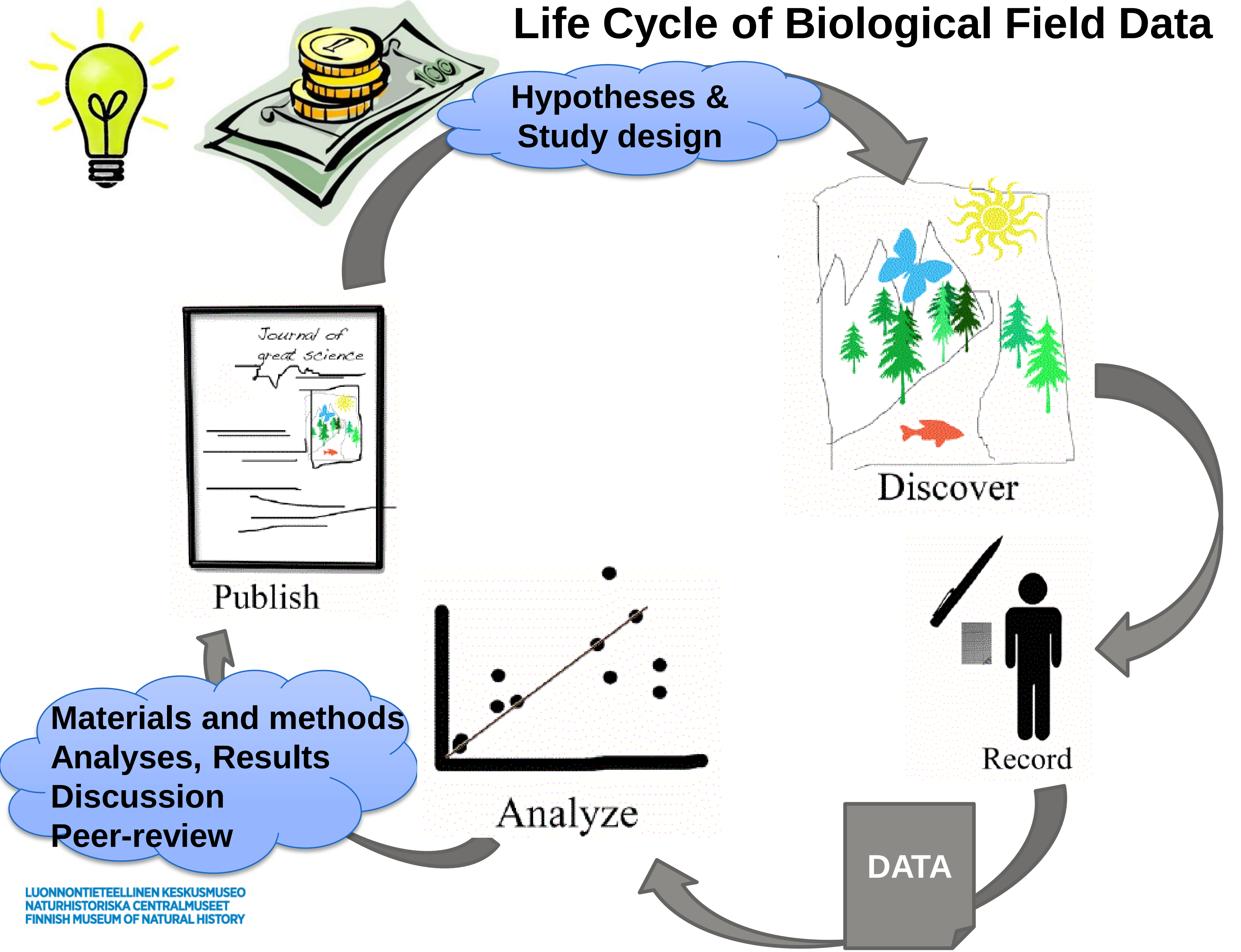
2016-11-02

What is Open Sciense?

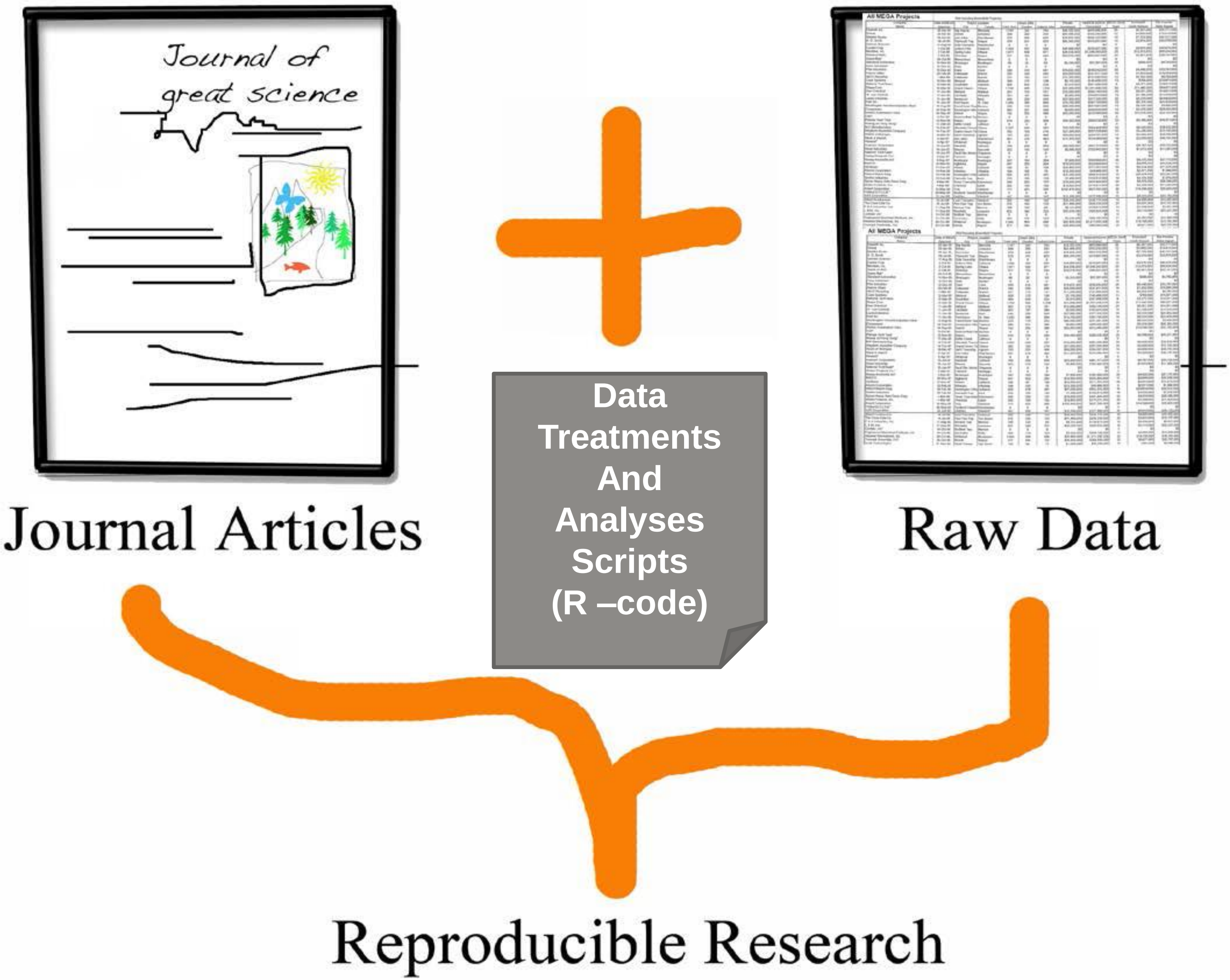


FOSTER
www.fosteropenscience.eu

Life Cycle of Biological Field Data



LUONNONTIETEELLINEN KESKUSMUSEO
NATURHISTORISKA CENTRALMUSEET
FINNISH MUSEUM OF NATURAL HISTORY


Journal of great science
All MEGA Projects
Journal Articles
Data Treatments And Analyses Scripts (R –code)
Raw Data
Reproducible Research

BENEFITS OF OPENESS

Increases the **efficiency** of research

Promotes scholarly rigor and enhances the **quality** of research

Enables tracking of **data use and data citation** through DOIs

Expands the spectrum of academic products through **data papers**

Enhances **visibility** and scope for engagement

Enables researchers to ask **new research questions**

Enhances **collaboration** and community-building

Increases the economic and social **impact of research**

Response to international conventions and **requirements from funding agencies**

OPEN ACCESS Freely available online

PLoS one

Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma

Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania, United States of America

Background. **Sharing research data provides benefit to the general scientific community, but the benefit is less obvious for the investigator who makes his or her data available.** ***Principal Findings.*** **We examined the citation history of 85 cancer microarray clinical trial publications with respect to the availability of their data. The 48% of trials with publicly available microarray data received 85% of the aggregate citations. Publicly available data was significantly (p = 0.006) associated with a 69% increase in citations, independently of journal impact factor, date of publication, and author country of origin using linear regression.** ***Significance.*** **This correlation between publicly available data and increased literature impact may further motivate investigators to share their detailed research data.**

Citation: Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

INTRODUCTION

Sharing information facilitates science. Publicly sharing detailed research data–sample attributes, clinical factors, patient outcomes, DNA sequences, raw mRNA microarray measurements–with other researchers allows these valuable resources to contribute far beyond their original analysis[1]. In addition to being used to confirm original results, raw data can be used to explore related or new hypotheses, particularly when combined with other publicly available data sets. Real data is indispensable when investigating and developing study methods, analysis techniques, and software implementations. The larger scientific community also benefits: sharing data encourages multiple perspectives, helps to identify errors, discourages fraud, is useful for training new researchers, and increases efficient use of funding and patient population resources by avoiding duplicate data collection.

Believing that that these benefits outweigh the costs of sharing research data, many initiatives actively encourage investigators to make their data available. Some journals, including the *PLoS* family, require the submission of detailed biomedical data to publicly available databases as a condition of publication[2–4]. Since 2003, the NIH has required a data sharing plan for all large funding grants. The growing open-access publishing movement will perhaps increase peer pressure to share data.

However, while the general research community benefits from shared data, much of the burden for sharing the data falls to the study investigator. Are there benefits for the investigators themselves?

A currency of value to many investigators is the number of times their publications are cited. Although limited as a proxy for the scientific contribution of a paper[5], citation counts are often used in research funding and promotion decisions and have even been assigned a salary-increase dollar value[6]. Boosting citation rate is

RESULTS

We studied the citations of 85 cancer microarray clinical trials published between January 1999 and April 2003, as identified in a systematic review by Ntzani and Ioannidis[7] and listed in Supplementary Text S1. We found 41 of the 85 clinical trials (48%) made their microarray data publicly available on the internet. Most data sets were located on lab websites (28), with a few found on publisher websites (4), or within public databases (6 in the Stanford Microarray Database (SMD)[8], 6 in Gene Expression Omnibus (GEO)[9], 2 in ArrayExpress[10], 2 in the NCI GeneExpression Data Portal (GEDP)(gedp.nci.nih.gov); some datasets in more than one location). The internet locations of the datasets are listed in Supplementary Text S2. The majority of datasets were made available concurrently with the trial publication, as illustrated within the WayBackMachine internet archives (www.archive.org/web/web.php) for 25 of the datasets and mention of supplementary data within the trial publication itself for 10 of the remaining 16 datasets. As seen in Table 1, trials published in high impact journals, prior to 2001, or with US authors were more likely to share their data.

The cohort of 85 trials was cited an aggregate of 6239 times in 2004–2005 by 3133 distinct articles (median of 1.0 cohort citation per article, range 1–23). The 48% of trials which shared their data received a total of 5334 citations (85% of aggregate), distributed as shown in Figure 1.

Academic Editor: John Ioannidis, University of Ioannina School of Medicine, Greece

Received December 13, 2006; **Accepted** February 26, 2007; **Published** March 21, 2007

Piwowar *et al.* (2007)

RAW DATA: original *unformatted* excel file or other original (machine produced) files

How to open these isolated data silos?

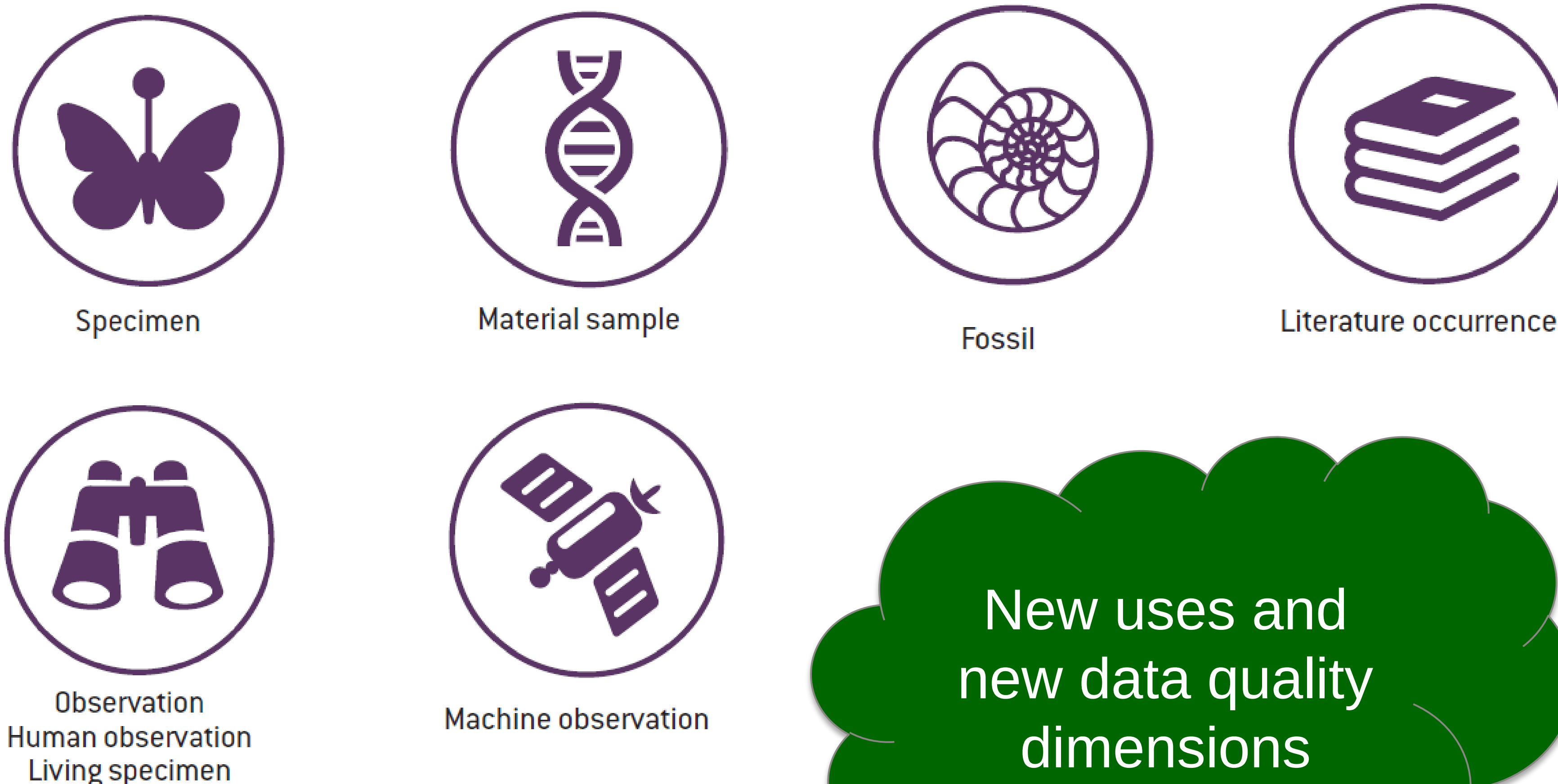
Core data types (in biodiversity data)

1. **Resource (or Dataset) Metadata**
   - Descriptive information about datasets
   - Metadata provides information about the **suppliers** of biodiversity data and about the **origins (provenance)**, **purpose** and nature of those data together with the statement of their '**fitness-for-use**'.

2. **Taxonomic Data**
   - Information relating to a taxon and NOT necessarily to a specific instance (occurrence)
   - Nomenclature and the taxonomical hierarchy of it, synonyms, type specimen categories, common names, historical names and checklists are taxonomic data

3. **Ocurrene Data (Primary Biodiversity Data)**
   - Occurrence of biological species in **spatial and temporal** terms is the fundamental data unit on which services and analytical workflows are based on. (*species-location-time*)

TYPES OF (OCCURRENCE) DATA SHARED THROUGH GBIF



New uses and new data quality dimensions

sample-based data

http://www.gbif.org/occurrence

What is a Standard?

- **Standards** are documented agreements on *representation, format, definition, structuring, tagging, transmission, manipulation, use, and management of data.*
- **A Standard provides a structure to describe data** with:
  - Common **terms** to allow *consistency* between records
  - Common **definitions** for easier *interpretation*
  - Common **language** for ease of *communication*
  - Common **structure** to quickly *locate* information
- In **search and retrieval**, standards provide:
  - Documentation structure in a reliable and predictable format for **computer interpretation**
  - A uniform summary description of the dataset

Standards make data interoperable with other data!

Biodiversity Information standards (TDWG)

http://www.tdwg.org/standards/

Metadata Standards:

- **Dublin Core => http://dublincore.org/**
- **Ecological Metadata Language EML (GBIF metadata profile)**
- **DataOne => https://www.dataone.org/education**

Standards for data exchange:

- **Darwin Core (based on Dublin Core, but adjusted for biodiversity data)**

- **Locality information standards**
  - OGC, GML (Geography Markup Language)
  - Gazeteers i.e. locality name databases

Darwin Core - a vocabulary of terms

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, De Giovanni R, Robertson T, and Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7(1): e29715. doi:10.1371/journal.pone.0029715

GBIF
www.gbif.org

http://rs.tdwg.org/dwc/terms/index.htm

Darwin Core Terms: A quick reference guide

Title:	Darwin Core Terms: A quick reference guide
Date Issued:	2009-02-12
Date Modified:	2015-06-02
Abstract:	This document is a quick reference for all recommended Darwin Core terms. For complete h changes and pre-standard terms, see [HISTORY]. For a comparative table of elements from current terms in the standard, see [VERSIONS].
Contributors:	John Wieczorek (MVZ), Markus Döring (GBIF), Renato De Giovanni (CRIA), Tim Robertson (G
Legal:	This document is governed by the standard legal, copyright, licensing provisions and disclai Group.
Part of TDWG Standard:	http://www.tdwg.org/standards/450/
Creator:	Darwin Core Task Group
Identifier:	http://rs.tdwg.org/dwc/2015-03-19/terms/
Latest Version:	http://rs.tdwg.org/dwc/terms/
Replaces:	http://rs.tdwg.org/dwc/2014-11-08/terms/
Document Status:	Current Standard

Taxonomic data

http://rs.tdwg.org/dwc/terms/index.htm#taxonindex

- **Reference to the checklist, synonymes, type categories**…
- 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

Term Name: taxonConceptID	
Identifier:	http://rs.tdwg.org/dwc/terms/taxonConceptID
Class:	http://rs.tdwg.org/dwc/terms/Taxon
Definition:	An identifier for the taxonomic concept to which the record refers - not for the nomenclatural details of a taxon.
Comment:	Example: "8fa58e08-08de-4ac1-b69c-1235340b7001". For discussion see http://terms.tdwg.org/wiki/dwc:taxonConceptID
Details:	taxonConceptID
Term Name: scientificName	
Identifier:	http://rs.tdwg.org/dwc/terms/scientificName
Class:	http://rs.tdwg.org/dwc/terms/Taxon
Definition:	The full scientific name, with authorship and date information if known. When forming part of an Identification, this should be the name in lowest level taxonomic can be determined. This term should not contain identification qualifications, which should instead be supplied in the IdentificationQualifier term.
Comment:	Examples: "Coleoptera" (order), "Vespertilionidae" (family), "Manis" (genus), "Ctenomys sociabilis" (genus + specificEpithet), "Ambystoma tigrinum diaboli" (genus + specificEpithet + infraspecificEpithet), "Roptrocerus typographi (Györfi, 1952)" (genus + specificEpithet + scientificNameAuthorship), "Quercus agrifolia var. oxyad (Torr.) J.T. Howell" (genus + specificEpithet + taxonRank + infraspecificEpithet + scientificNameAuthorship). For discussion see http://terms.tdwg.org/wiki/dwc:scientificName

Ocurrence data

http://rs.tdwg.org/dwc/terms/index.htm#occurrenceindex

- **DarwinCore (DwC)** and **Access to Biological Collection Data (ABCD)**
- Specimen identification or absence, associated species, locality, time (frame) and measurements or facts (for the gathering event)

- New feature in IPT (GBIF Integrated Publishing Toolkit) and DwC supports sharing sample based data by describing "**Events**" within a dataset (event metadata)
- More detailed technical information and discussion http://terms.tdwg.org/wiki/dwc:samplingProtocol

Event | **HumanObservation** | **MachineObservation**

eventID | parentEventID | fieldNumber | eventDate | eventTime | startDayOfYear | endDayOfYear | year | month | day | verbatimEventDate | habit
sampleSizeUnit | samplingEffort | fieldNotes | eventRemarks

Location

locationID | higherGeographyID | higherGeography | continent | waterBody | islandGroup | island | country | countryCode | stateProvince | county |
minimumElevationInMeters | maximumElevationInMeters | verbatimElevation | minimumDepthInMeters | maximumDepthInMeters | verbatimDepth | mini
maximumDistanceAboveSurfaceInMeters | locationAccordingTo | locationRemarks | decimalLatitude | decimalLongitude | geodeticDatum | coordinateU
pointRadiusSpatialFit | verbatimCoordinates | verbatimLatitude | verbatimLongitude | verbatimCoordinateSystem | verbatimSRS | footprintWKT | foot
georeferencedBy | georeferencedDate | georeferenceProtocol | georeferenceSources | georeferenceVerificationStatus | georeferenceRemarks

GeologicalContext

geologicalContextID | earliestEonOrLowestEonothem | latestEonOrHighestEonothem | earliestEraOrLowestErathem | latestEraOrHighestErathem | earli
latestPeriodOrHighestSystem | earliestEpochOrLowestSeries | latestEpochOrHighestSeries | earliestAgeOrLowestStage | latestAgeOrHighestStage | lo
highestBiostratigraphicZone | lithostratigraphicTerms | group | formation | member | bed

Identification

identificationID | identificationQualifier | typeStatus | identifiedBy | dateIdentified | identificationReferences | identificationVerificationStatus | identi

SCIENCE INSTRUMENTS

Existing:

- Research planning
- Collecting data
- Data analysis
- Publication
- Distribution

Data

... collecting

... input

... editing

... analysis

... archival

Often overlooked:

- Data management

SHORT TERM → **perspective** → **LONG TERM**

RESEARCH PHASE

- file formats
- ownership
- metadata
- storage
- backups

DISSEMINATION PHASE

- share with whom?
- embargo?
- licensing
- metadata

PRESERVATION PHASE

- repository?
- long-term manager?

BARRIERS AND CONSTRAINTS TO OPENNESS

- Lack of evidence of benefits and rewards
- Lack of system demand
- Lack of skills, time and other resources
- Cultures of independence and competition
- Concerns about quality
- Ethical, legal and other restrictions on accessibility

Open to All?
Case studies of openness in research

A joint RIN/NESTA report | September 2010

Data Quality (DQ) and Quality Assurance (QA)

Data Validation – the importance of planning for and creating tidy, standardized data

- **Avoiding errors already in the field**
- **Avoiding systematic errors when correcting**
- **Documentation (make correcting or re-identification possible)**
- **Recording accuracy and uncertainty whenever possible!!**

Loss of data quality can occur at many stages:

- At the time of collection
- During digitisation
- During documentation
- During storage and archiving
- During analysis and manipulation
- At time of presentation
- And through the use to which they are put

In general, error must not be treated as a potentially embarrassing inconvenience, because error provides a critical component in judging fitness for use.

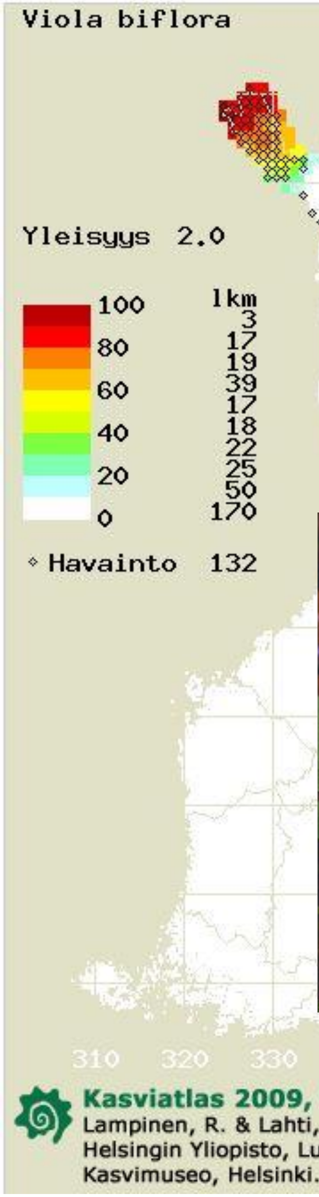
Chrisman, 1991

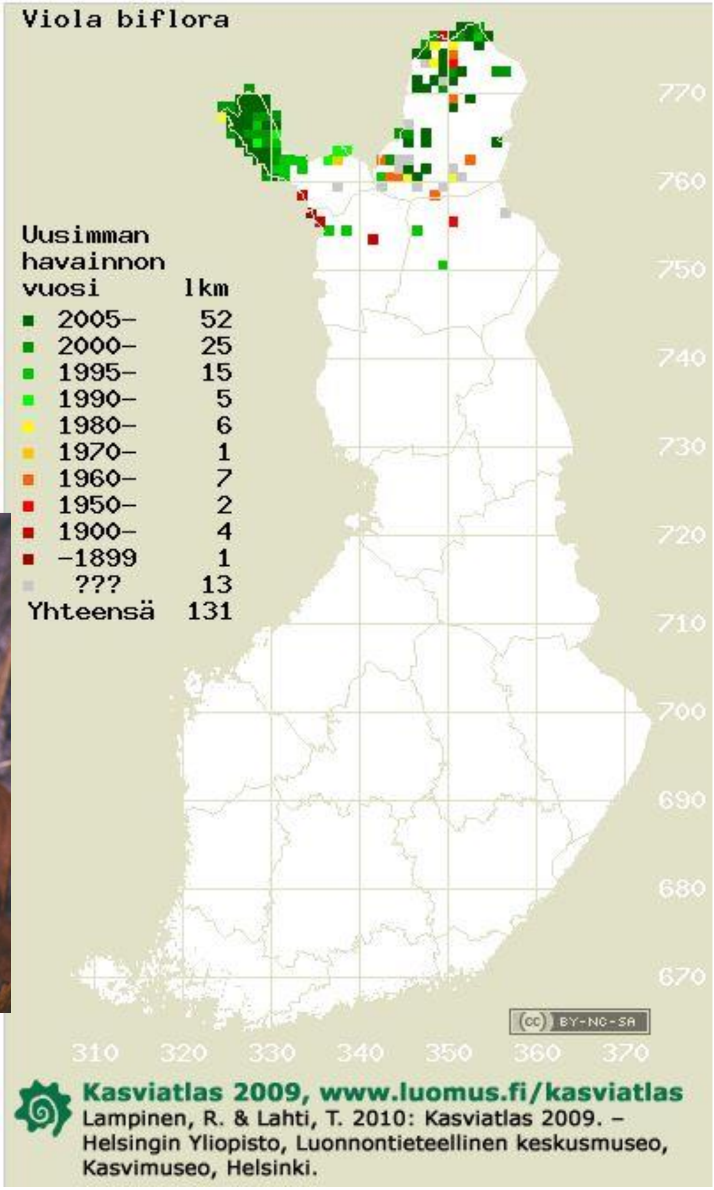
Fitness-For-Use means that the quality of the data has been documented so, that the user can estimate whether the data is fit to be used for his/her purposes

Using Open Species Data

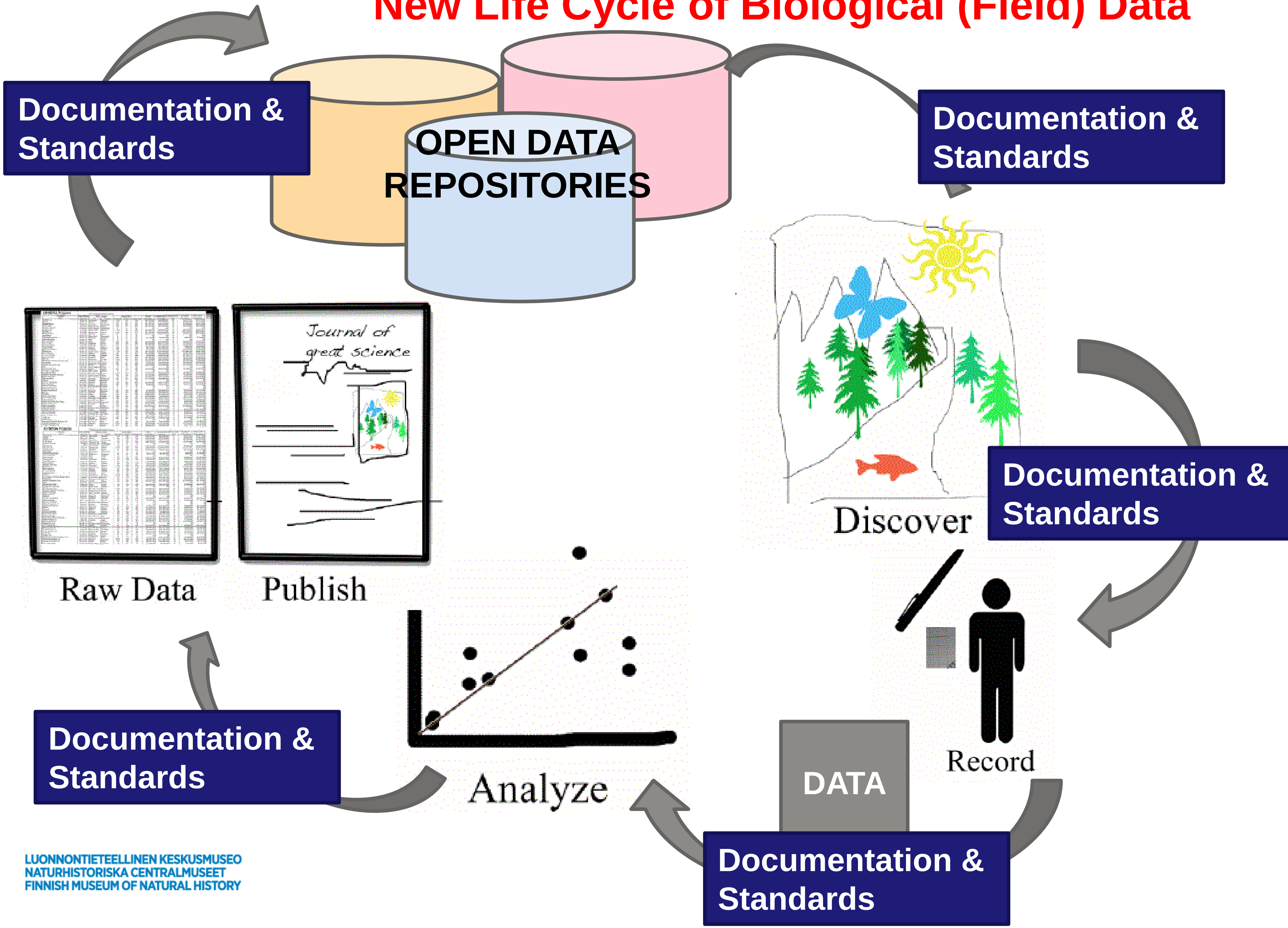
- Biogeographic Studies, Species Modelling
- Species Diversity and Population studies
- Life Histories and Phenologies
- Studies of Threatened and Migratory species
- Climate Change Impacts
- Ecology, Ecosystems, Evolution and Genetics
- Environmental Regionalisations
- Conservation Planning
- Natural Resource Management

Viola biflora, Twoflower Violet, Arctic Wood Violet, Arctic Yellow Violet





New Life Cycle of Biological (Field) Data



LUONNONTIETEELLINEN KESKUSMUSEO
NATURHISTORISKA CENTRALMUSEET
FINNISH MUSEUM OF NATURAL HISTORY

References:

- TDWG – Biodiversity Information Standards http://rs.tdwg.org/dwc/index.htm
- DataOne > Education modules https://www.dataone.org/education
- Chapman, A. D. 2005. **Principles of Data Quality**, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. ISBN 87-92020-03-8. Available online at http://www.gbif.org/orc/?doc_id=1229.
- Chapman, A. D. 2005. **Principles of Data Cleaning – Primary Species and Species Occurrene Data**, version 1.0. Report for the Global Biodiversity Information Facility, Copenhagen. Available online at: http://www.gbif.org/resource/80528