Biodiversity Informatics

Introduction and overview of landscape

Dr Dimitris Koureas

Biodiversity Informatics | Department of Life Sciences
Natural History Museum London

Research Data Alliance Biodiversity Data Integration IG - Chair
Biodiversity Information Standards – TDWG - Exec. member
The four parts of this presentation

1. Background and vision for biodiversity Informatics
2. Challenges and approach
3. Key actors in the landscape
4. Technical discussion
Biodiversity informatics

The study of the **transformation** and **communication** of **information** in Life and Earth sciences

provides the **means**
(generating and enhancing the necessary infrastructure)
Why Biodiversity Informatics is now more important than ever before?
The problem – integrating biodiversity research

**How to we join up these activities?**
What infrastructures do we need? (technologies, tools, standards...)
What processes do we need? (Modelling, workflows...)
What data do we need? (Genes, localities...)

**How do we use this as a tool?**
Species conservation & protected areas
Impacts of human development
Biodiversity & human health
Impacts of climate change
Food, farming & biofuels
Invasive alien species
Data is everywhere and is produced with an ever increasing rate

90% of all science data generated in the last 3 years!
Data, Data, Everywhere, Nor Any Drop to Drink

Christine L. Borgman
Professor and Presidential Chair in Information Studies
University of California, Los Angeles

Keynote presentation
Research Data Alliance, Fourth Plenary Meeting
Amsterdam, September 2014

Gustave Dore, *Rime of the Ancient Mariner*, Woodcut, 1798

Share, Structure, Describe, Aggregate and Preserve data
A global shift in the modus operandi of Science
Open Data-driven Science

EUROPEAN COMMISSION
DIRECTORATES-GENERAL FOR RESEARCH AND INNOVATION (RTD) AND COMMUNICATIONS NETWORKS, CONTENT AND TECHNOLOGY (CONNECT)

BACKGROUND DOCUMENT

PUBLIC CONSULTATION
"SCIENCE 2.0": SCIENCE IN TRANSITION
What are the **challenges** for **Biodiversity Informatics** today?
Publications based on countless specimens, images, maps, keys and datasets typically generated by small communities for "local" research projects.
Current taxonomic data production

- **15-20k** new spp. described annually (2M total)\(^1\)
- **30k** nomenclatural acts (12M total) \(^1\)
- **20k** phylogenies (750k total) \(^2\)
- **31k** taxa sequenced (360k taxa total) \(^3\)
- **800k** BioMed papers (40M total pp. of taxonomy) \(^4\)

- Countless **specimens, images, maps, keys and datasets**

1.8 M described spp. (20M names)

300M pages (over last 250 years)

1.5-3B specimens

Now imagine that...

Estimates of

7.5 million species
still undescribed

$y = 0.46x^{1.03}$  
$R^2 = 0.96$

1How Many Species Are There on Earth and in the Ocean? Mora C et al.  
doi:10.1371/journal.pbio.1001127
Expected volume of *taxonomic* and *biodiversity* data

Need of *extracting*, *aggregating* and *linking* data on a global level
Science is **global**
- It needs global standards
- Global workflows
- Cooperation of global players

**BUT**

Science is carried out **“locally”**
- By local scientists
- Being part of local infrastructures
- Having local funders
Mobilising the long tail of data

High visibility

Inaccessible Dark data
“Link together evolutionary data... by developing analytical tools and proper documentation and then use this framework to conduct comparative analyses, studies of evolutionary process and biodiversity analyses”

This requires data, information & knowledge to be...

- Digital  
  Not printed paper
- Openly accessible  
  Not behind barriers (e.g. paywalls)
- Linked-up  
  Not in silos
Research vs Infrastructure

Slide adapted from Patterson D. 2013, Tempe, Arizona
Research

- Discovery
- Ephemeral
- Individualistic
- Massive redundancy
- Optional
- Risk taking

vs

Infrastructure

Slide adapted from Patterson D. 2013, Tempe, Arizona
Research

- Discovery
- Ephemeral
- Individualistic
- Massive redundancy
- Optional
- Risk taking

vs

- Implementation
- Communal / agreed
- Essential
- Persistent
- Robust & reliable
- Adaptable

Infrastructure

Slide adapted from Patterson D. 2013, Tempe, Arizona
Key problems
• Landscape is complex, fragmented & hard to navigate
• Many audiences (policy makers, scientists, amateurs, citizen scientists)
• Many scales (global solutions to local problems)

Figure adapted from Peterson et al, Syst. & Biodiv. 2010
doi: 10.1080/14772001003739369
Hour-glass motif for big data infrastructure

Communities

Communities

Data pool

Data generation

Data re-use

Slide adapted from Patterson D. 2013, Tempe, Arizona
Big data world with re-use data

- Re-use
- Quality enhancement

- Distribute
- Make discoverable and actionable
- Atomize
- Standardize (metadata, ontology)
- Use stable UUIDs to identify content

- Preserve
- Federate
- Register

- Make accessible
- Normalize data
- Structure data
- Make data digital

Data pool

Visualization → Analysis → Aggregation → Manipulation

Data generation → Data re-use
Big data world with re-use data

Visualization  Analysis  Aggregation  Manipulation

Data re-use

Data generation

Observations  Experiments  Models  Processed
Nodes interconnected

- Dynamically interconnected
- Nodes with sub-discipline specific responsibilities
- **Standard Exchange formats**
- Using UUIDs to identify content
- Ontologies

Nodes are the essence of infrastructure

Slide adapted from Patterson D. 2013, Tempe, Arizona
But... how many biodiversity informatics projects and tools are out there?
But... how many biodiversity informatics projects and tools are out there?

At least 679!

Categories:

**Data Aggregator** - a web site that collates data from a variety of sources (digital and hardcopy) and presents it in one form

**Data Indexer** - a web site that provides lists or indexes of other sites that provide data

**Data Provider** - a web site that provides data directly from research or other studies

**Data Standards** - a web site that contributes to formulating or developing standards for data

**Facilitator** - a web site that facilitates the provision of data by other projects or web sites

Sources: EDIT, TDWG & ViBRANT
Occurrence data aggregated from different nodes (data holders)
Encyclopedia of Life
http://eol.org

Global access to knowledge about life on Earth
EOL - TraitBank

Over 8 million traits
Species+
http://www.speciesplus.net/

A combined source for legislation, distribution and trade in MEA-listed species
Making taxonomy **digital, open & linked**

Scratchpads

biodiversity online

http://scratchpads.eu
Making taxonomy **digital, open & linked**

Scratchpads
biodiversity online
http://scratchpads.eu
The Scratchpads concept

Your data

Scratchpads
biodiversity online

External data & services

- eMonocot
- eOL
- GBIF
- NCBI
- WoRMS
- Google Scholar
- BHL
- Biodiversity Heritage Library
650 Scratchpads Communities

by 6,570 active registered users

covering 178,800 taxa

in 1,200,000 pages.

In total more than 1,900,000 visitors

70,000 unique visitors/month
iPlant collaborative
http://www.iplantcollaborative.org/

Platforms and Tools

Discovery Environment
- Use hundreds of bioinformatics Apps and manage data in a simple web interface.

DNA Subway
- Take DNA Subway to teach classroom-friendly bioinformatics for genome analysis, DNA Barcoding, and RNA-Seq.

Bisque Image Analysis Environment
- Exchange, explore, and analyze biological images and their metadata.

Atmosphere
- Create a custom cloud-based scientific analysis platform or use a ready-made one for your area of scientific interest.

Data Store
- Store, manage, access, and share all the data related to your research.

User Portal

About iPlant

The Project
- Building cyberinfrastructure for life sciences research.

Science
- Enabling discovery in the age of Big Data.

Success Stories
- How users and projects have used iPlant.

Publications
- Recent publications and how to cite iPlant.

Help
- Find answers to common questions.

Learning Center

Get Started
- First things to do when you get your account.

Platform and Science Tutorials
- Step-by-step guides for our platforms and specific analyses.

Ask iPlant
- Post your science and support questions.
GenBank

A comprehensive database that contains publicly available **nucleotide sequences** for almost **260,000 formally described species**

GenBank is part of the International Nucleotide Sequence Database Collaboration
Researchers can assemble, test, and analyse their data records in BOLD before uploading them to: *International Nucleotide Sequence Database Collaboration* (DDBJ, ENA, GenBank)
Catalogue of Life
http://catalogueoflife.org

A single authoritative source of taxonomic information
Biodiversity Heritage Library (BHL)
http://www.biodiversitylibrary.org/

Biodiversity literature openly available to the world

> 200M pages of legacy literature
Data aggregation and application of informatics tools in the study of biodiversity is associated with major challenges.
Data exchange standards

Darwin Core (DwC) Primarily used as a specimen records metadata standard
http://rs.tdwg.org/dwc/index.htm

Access to Biological Collection Data (ABCD)
http://www.tdwg.org/standards/115/

Audubon Core Multimedia Resources Metadata Schema
http://www.tdwg.org/standards/638/
Standards facilitate systems interoperability
We need Unique Identifiers

Identifiers
A key to find something in a database.

UPIDs to identify content
We need Unique Identifiers

10.4289/0013-8797.115.1.75
We need Unique Identifiers

http://hdl.handle.net/10.4289/0013-8797.115.1.75

http://dx.doi.org/10.4289/0013-8797.115.1.75

http://www.google.co.uk/search?q=10.4289/0013-8797.115.1.75

http://zoobank.org/10.4289/0013-8797.115.1.75
We need Unique Identifiers

Can a taxonomic name be used as a UPID?

Are taxonomic names enough for communication between Scientists? **YES**

Are taxonomic names enough for communication between machines?

Is it **Unique**?
Is it **Persistent**?
Is it an **Identifier**?
We need Unique Identifiers (e.g. think of homonyms)

genus *Satyrium*
We need Unique Identifiers

For example:

Figure 3: A search of GenBank for Melissotarsus insularis finds no sequences. However, a search of AntWeb finds a specimen listed as having been barcoded, and the paper publishing the barcodes has a supplementary table that lists the specimen as the source for sequence DQ176312. GenBank lists this sequence as being from taxon Melissotarsus sp. BLF ml.
Ontologies

Knowledge Organisation Systems

The need for **Controlled Vocabularies** and **Ontologies**

*Google has done it:*
http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html

**Plant anatomical and structural development Ontology**
http://www.obofoundry.org
http://www.plantontology.org/
Example of ontology usage

Deans A. et al. Time to change how we describe biodiversity, Trends in Ecology & Evolution 2012
doi:10.1016/j.tree.2011.11.007
Biodiversity scientists need access to tools and services

Provision of these tools and services requires efficient e-infrastructures

E-Infrastructure is predicated on the use of a agreed upon technical backbone
OLD JOKE

A drunk is crawling around a lamp post on his hands and knees.
A cop comes along ...

**Cop:** What are you doing?

**Drunk:** Looking for my car keys.

**Cop:** Are you sure you dropped them here?

**Drunk:** No, I dropped them in the alley.

**Cop:** So why are you looking here?

**Drunk:** Because the light’s better.
Science is a ‘light’s better’ endeavour in that research effort is not directed at areas where the work is technically infeasible. Research is directed where real, interpretable results may be obtained.

We do, in fact, conduct research where the light’s better. But, when the light changes, so does science. With better illumination, we look in new areas. We find new things...
Thank you

Dimitris Koureas

d.koureas@nhm.ac.uk | @DimitrisKoureas

http://mvz.berkeley.edu/Informatics.html
Data

For example
Image
Specimen
Dataset (e.g. database/excel)

Metadata

For example
Photographer, ISO settings
Collector, locality
Description of columns/fields