Digital Collections – towards a UK & European vision

Vince Smith, NHM London

Bristol Workshop - Towards a strategic approach to mobilising UK museum biodiversity data, M-Shed, UK

8 March 2018
Natural Science Collections

Bio- & geo-diversity
Leading scientific facilities
Engaging the public
Discovering, describing and interpreting life on Earth
Tackling societal challenges
Sustainable future
Molecular repository
NHM Digital Collections Programme
(2014-2024)

“To collate, organise and make available to global scientific & public audiences one of the world’s most important natural history collections”

‘Ambition’ to digitise 20 million specimens
<table>
<thead>
<tr>
<th>POLICY &amp; PROTOCOL</th>
<th>2 Year</th>
<th>5 Year</th>
<th>10 Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defined data policy &amp; standards</td>
<td>Policies embedded in operating practices</td>
<td>World leaders in digital curation</td>
<td></td>
</tr>
<tr>
<td>DATA CAPTURE</td>
<td>Prioritised digitisation workflows piloted</td>
<td>Portfolio of mass digitisation projects</td>
<td>Some major collections digitised</td>
</tr>
<tr>
<td>PEOPLE &amp; SKILLS</td>
<td>Task force formed and operating</td>
<td>Best practice processes integrated into training</td>
<td>Digital curation core part of BAU practice</td>
</tr>
<tr>
<td>INFRASTRUCTURE</td>
<td>Refined collections database, tools &amp; hardware</td>
<td>Future collections database implemented</td>
<td>Broad connections to other large digital collections</td>
</tr>
<tr>
<td>STAKEHOLDERS &amp; GOVERNANCE</td>
<td>Key user communities engaged</td>
<td>Peer-to-peer development</td>
<td>Proactive engagement of emerging audiences</td>
</tr>
<tr>
<td>PARTNERSHIPS</td>
<td>Partners involved in pilot projects</td>
<td>Fully-funded digitisation portfolio</td>
<td>Major international coalitions</td>
</tr>
<tr>
<td>RESEARCH</td>
<td>Research-orientated projects and initiatives</td>
<td>Collaborative research material published</td>
<td>Major contributions to grand challenges</td>
</tr>
<tr>
<td>ACCESS</td>
<td>Live NHM Data Portal</td>
<td>Tools, visualisations &amp; analytics</td>
<td>Integrated global network of users</td>
</tr>
</tbody>
</table>

**NHM Digital Programme**

- 2, 5 & 10 yr. phases
- A change programme
- Addresses historic under investment
- Multiple digitisation projects
- Focus on innovation & science

**NHM Digital Programme**
Pilots

Crowdsourcing

Micro Collections

Informatics

Lepidoptera

Acoustics & 3D

Open Herbarium
Environmental Change
- UK butterflies & moths
- 800k specimens
- 2 mins per specimen
- £1 per specimen

Large-scale digitisation:
- High-throughput digitisation workflows
- Informatic pipelines
- Computer-assisted object recognition
Parasites & Vectors:
- High-throughput microscope slide digitisation
- Low cost primary setup, high throughput
- A second attempt at slides (low tech.)

Parasitic lice (new spp. & cospeciation):
- Circa 70-80k slides (20k to date)
- 600-1k slides per day pp. (£0.07-0.2 ea.)
- Twin tracks (low & high resolution)
- Setup, imaging, transcription, Q.C. barcoding, databased (but not research)
Data Portal

- NHM research & collections data
- Access, reuse & citation
- 9.1m records, 95 datasets since 2015
- Images, sound, video & 3D
- Default open licensing
- GBIF
- 10B records downloaded in 139k events, 61 papers
Impact

- Publications citing NHM data (GBIF only):
  - Open access: 59
  - Peer-reviewed: 50
  - Open Access: 22

GBIF topic tags by frequency:

- Evolution Ecology
- Phylogenetics
- Ecosystem Services
- Invasive Ecology
- Conservation
- Climate Change
- Biodiversity Science
- Species Distributions
- Taxonomy Evolution Ecology

Altmetrics:
- Collection specimens: 109
- Collection specimens: 95
- Collection specimens: 42

*Altmetrics are non-traditional metrics used to measure the impact of a scholarly work. The "donuts" above mainly reflect tweets referencing NHM data DCC.
Challenges

1. Mass Digitisation
   Specimen image

2. Data Extraction
   - Transcription (e.g., Zooniverse)
   - OCR (e.g., reCAPTCHA)
   - Georeferencing (Online mapping tools)
   - Image Recognition (e.g., Image Search)

3. Using the Data
   - Linking to Archives and Literature (e.g., Biodiversity Heritage Library)
   - Analytical Tools (e.g., OpenRefine)
   - Data Visualisation (e.g., CartoDB)
   - Search (e.g., WikiData)

4. End Products
   - NHM in an App
   - Digital Exhibitions (Virtual Tours)
Environmental Change

Measuring biodiversity loss

Digitisation workflows

Measuring hyper diversity

Data visualisation

Species recognition

Computer vision
Database (biodiversity, land use & human population)

Projection (predicted change to species richness)

Model (Biodiversity intactness Index – BII)

Newbold et al. 2016
Science 353:288-291
55% of the world’s collections with rich historical & global distribution

- **1.5 billion** specimens
- **80%** of world’s species
- **5,000** scientists employed
- **16,000** scientific visitors pa
- **10 million** public visitors pa
- **25 million** web visitors pa
Projects

Major digital initiatives

SYNTHESES
Synthesis of systematic resources

ViBRANT
Virtual Biodiversity

eMonocot

Alignment with organisational agendas

Major funders

European Commission

NRC
Science of the Environment

Major collaborators

Catalogue of Life

Biodiversity Information Standards
CETAF
RDA

GIF
GBIF
EOL
GGBN

• 13-year programme of physical & digital access
• SYNTHESES1,2,3 (>50k access days, 3.8k sci., 4.5k pubs.)
• Focus on science, policy and public engage. activities
• A unique role within the infrastructure landscape
• Foundation for deeper integration
DiSSCo: A new European infrastructure

114 National Facilities
21 Countries

- Largest ever formal agreement between natural science collection facilities
- Centralised governance model already in place
- Supporting network of working groups

With political support by 11 European Governments
1. **e-Science services**
   - A one-stop shop for services providing unified discovery, access, interpretation and analysis of complex linked data.

2. **Physical and remote access services**
   - A universal harmonised physical access service and digitisation on demand service. Supporting mass-scale, yet synchronised content mobilisation across European Collections.

3. **Support & Training services**
   - Integrated user support desk and implementation of multi-modal training programmes to enhance data skills.
• Support a coherent approach to policy on European research infrastructures (RI’s)

• Facilitate initiatives leading to better use & development of RI’s

• Complex, two-year roadmap admission process

• Two delegates per country, nominated by the national research/science agencies.
Costs

Preparatory & Construction
€89.6M

Operational
€12.1M / year
(plus additional digitisation investments)

Resourcing

EC funding
Consortium investments

EFSI
Consortium (in-kind)
Member State contributions

H2020 (CSA)
COST Action
Consortium (in-kind)
Member State contributions

H2020 (RIA)
Public Private Partnerships
Consortium (in-kind)
Member State contributions

European Investment Bank
EFISI
Consortium (in-kind)
Member State subscriptions
An urgently needed initiative

- Official ESFRI decision due June 2018
- If successful, DiSSCo is open to new organisations
  - Opportunity for leadership at a global level
  - Direct response to identified needs in the European and international RI landscape
  - Lowers the barrier for big, open science practices across tens of thousands of users

Potential to improve the case for support for UK collections?
Potential to increase UK coordination on:
- Regional collections
- Observer networks
- Technical support
- Citizen science
- Education
- Others?

What might the “One-Collection” vision mean for the UK?
Can we be more than the sum of our parts?