No specimen left behind:
The NHM Digital Collection Programme

Vince Smith, Natural History Museum, London
Digitisation of Biological Collections, 13 April 2014
1.5-3 BILLION SPECIMENS
1.9 million species
300 years of collection
Digital surrogates of specimens:
- 3D models and printing
- CT imaging of internal structures
- Atomic, physical and chemical analysis
Using digitisation to enhance public engagement with natural history specimens:

- Augmented reality in gallery
- Animated content for pre- and post-visit apps
An inordinate fondness for beetles: 400,000 species described so far, 100,000 types among the NHM’s 10M beetle specimens.
### NHM collection

<table>
<thead>
<tr>
<th>Collection area</th>
<th>No of objects</th>
<th>No of type specimens</th>
<th>Physical register</th>
<th>Digital data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palaeontology</td>
<td>6,919,207</td>
<td>43,146</td>
<td>2,364,232</td>
<td>340,636</td>
</tr>
<tr>
<td>Mineralogy</td>
<td>423,563</td>
<td>615</td>
<td>425,000</td>
<td>402,727</td>
</tr>
<tr>
<td>Botany</td>
<td>5,863,000</td>
<td>172,750</td>
<td>127,200</td>
<td>645,222</td>
</tr>
<tr>
<td>Entomology</td>
<td>33,753,257</td>
<td>612,796</td>
<td>57,197</td>
<td>255,000</td>
</tr>
<tr>
<td>Zoology</td>
<td>27,501,350</td>
<td>325,000</td>
<td>1,986,000</td>
<td>1,160,216</td>
</tr>
<tr>
<td>Library &amp; archives</td>
<td>5,460,000</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>79,920,377</strong></td>
<td><strong>1,154,307</strong></td>
<td><strong>4,959,629</strong></td>
<td><strong>2,803,801</strong></td>
</tr>
</tbody>
</table>

<3% of NHM specimens are digitised, & even fewer are ‘computable’
Digital NHM: a strategic priority

Museum Strategy 2015-2020

Thee big narratives
① Origins and evolution
② Living diversity
③ Sustainable futures

Four priority areas
① Digital
② National
③ International
④ London

<table>
<thead>
<tr>
<th>Priority Area #1: Digital</th>
<th>Objectives</th>
<th>Key projects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Big, open data</strong></td>
<td>Mass collection digitisation*&lt;br&gt;NHM Data portal*&lt;br&gt;Crowd-sourcing of data*</td>
<td></td>
</tr>
<tr>
<td><strong>Global communities</strong></td>
<td>New website*&lt;br&gt;Social media channels Membership</td>
<td></td>
</tr>
<tr>
<td><strong>Innovative platforms</strong></td>
<td>Location-aware Wi-Fi*&lt;br&gt;Museum App*&lt;br&gt;In-gallery technology</td>
<td></td>
</tr>
</tbody>
</table>
iCollections: 2013-2015

- A pilot for the Digital Collections Programme
- iCollections digitisation criteria:
  - Entire collection
  - No existing digitisation pipeline
    - (pinned, slide & herbarium specimens)
  - High research potential
    - (phenology, morphometrics, migration patterns, pest associations, automated species recognition)
  - Curation opportunity

- Data outputs
  - Image(s)
  - Metadata (what, when & where)
  - Georeference point localities
Large-scale digitisation:
- High-throughput digitisation workflows
- Informatic pipelines
- Computer-assisted object recognition

NHM iCollections project:
- UK butterflies & moths
- 500k specimens
- 2 mins per specimen
- £1 per specimen
iCollections digitisation process
• Many species emerging earlier and earlier each year
• Initial collection date & temperature highly correlated
• A unique marker on phenological response before recent climate change
• Longer time perspective than most observational records (BMS post-1976)
• Museum data available for rare or hard to record species

Cooler springs

Warmer springs

Orange tip butterfly (Anthocharis cardamines)
NHM Digital Collections Programme

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

- an online specimen- / lot-level data base for all holding
- core meta-data and / or images for key parts of the collection, and
- flexible informatics and visualisation tools”

Target = 20 million specimens digitised in 5 years
PEOPLE & SKILLS
Prioritised digitisation
Workflows piloted
Portfolio of mass digitisation output projects
Some major collections digitised

DATA CAPTURE
Defined data policy and standards
Policies embedded in NHM operating practices
Leaders of process in the digital curatorial world

PEOPLE & SKILLS
Task force formed and operating
Best-practice processes integrated into training
Digital curation as a core part of our practice

INFRASTRUCTURE
Refined collections database, tools & hardware
Future collections database implemented
Broad connections to other large digital collections

STAKEHOLDERS & GOVERNANCE
Key user communities engaged
Peer to peer development
Proactive engagement of emerging audiences

PARTNERSHIPS
Partners involved in pilot projects
Fully funded digitisation portfolio
Major international coalitions

RESEARCH
Research-orientated projects & initiatives
Collaborative research material published
Major contributions to grand challenges

ACCESS
Live NHM Data Portal
Tools, visualisations & analytics
Integrated global network of users
Digital Collections Task Force

- Core pool of expertise, ideas and capacity for the Museum’s digitisation activity
- Key individuals across departments
- 50% or more of their time on projects under DCP
- Individuals seconded to relevant projects
- Coordinated by a steering group made up of units lead individuals and chaired by the Head of Collections Digitisation
- Membership expected to change as the portfolio of projects evolves
- Change from within
What meta-data are required for different types of study?
## Digital Collections Programme: pilot projects

### Pilot 1 – Herbarium sheets
- Trial partnership with Kew & Picturae (cf Naturalis)
- Conveyor-based imaging equipment
- Outsourcing of label transcription
- C. 70,000 sheets
- January-April 2015

### Pilot 2 – Palaeontology
- British Fossils (Mesozoic vertebrates)
- Standardised photography
- Capturing taxonomic & stratigraphic metadata
- February 2015 – March 2016

### Pilot 3 – Microscopic slides
- High-throughput slide scanner
- SatScan to capture label information
- Call for pilot project ideas
- April 2015 onwards
NHM & Kew Herbarium sheet digitisation

33k Specimens per day, 3 shifts (6am-10pm), Netherlands collection complete in 1.5 years
€1.29 Euros per specimen image (if outsourced), transcription at similar cost
Crowdsourcing
Crowdsourcing platforms

Herbarium @Home
http://herbariaunited.org/atHome/

Smithsonian Transcription Center
https://transcription.si.edu/

Atlas of Living Australia
http://volunteer.ala.org.au/

Les Herbonautes
http://lesherbonautes.mnhn.fr/

Notes from Nature
http://www.notesfromnature

Crowdcrafting
http://crowdcrafting.org

Research *by the people, for the people.*

We have hundreds of projects waiting for your help to achieve amazing goals.
NHM Ornithology Registers (1837 – 1990)

Progress
• Total Images: 2,950
• Active Images: 126
• Complete Images: 2,824
• 307,305 transcriptions
• Circa 1/3rd by one person
Science uncovered 2014: “Crowdsourcing the Collection”

- Demonstrate digitisation process
- Engage the public in transcription
  - Digitise; web publish images; public transcription; data publication
- Dedicated mobile website
NHM Data Portal

- A platform for deposition and discovery of NHM research & collections data
- Weekly updates from NHM collection database
- Stable, citable (DataCite) identifiers on datasets & records
- Transparent data quality indicators
- Export, import, & web-services available (Linked Open Data coming soon)
- Built using CKAN, with enhanced mapping, statistical and filtering functions
NHM Open Data Policy

Open-by-default policy on collections data & research data sets
- Creative Commons (CC) Zero waiver on data
- CC BY 4.0 licence on images

Exceptions to default
- Exceptions require justification, which include:
  - Commercial value
  - Research competitiveness
  - Third party rights
  - Sensitive collection information
  - Donor or funder conditions
  - Confidential documents

PSI Directive
- EU legislation on reuse of public sector information, extended to cover museums, libraries and public archives. To be implemented in July 2015
- General principal that information should be available for re-use for free or at marginal cost, and if reused, made available to all under equal terms.