Two part presentation:

1. The EUDAT – Herbadrop project.
   Lifted slide by slide from Pascal Dugénie, Herbadrop Project Officer - C.I.N.E.S (Centre Informatique National de l'Enseignement Supérieur)

2. The reality of being a data partner with HERBADROP
   Robert Cubey – Plant Records Officer – RBGE (Royal Botanic Garden Edinburgh)
The Herbadrop Data Pilot

- EUDAT Data Pilot
- Start date: Jan. 2016
- 18 months
Herbarium Images

• Millions of specimens in herbaria all over the world
• Global trend to industrial digitizing
• Data difficult to handle even for medium size institutes
• Same challenges being faced by potentially hundreds of herbaria in Europe
• Makes sense to work together to develop a solution
Herbadrop challenges

- Specimen images available online
  - cross institute & domains data analyses
- Large datasets of herbarium specimens
  - 500 million worldwide
  - 22 million in Germany
  - 20 million in France
- Extracting information from the specimen labels:
  - enhanced methods using state-of-the-art OCR technologies
  - particularly complex due to the presence of biological material in the image
- Discovering new links between objects of different collections
Principles of OCR processing
Mining for duplicates
Herbadrop objectives

1. **PRESERVATION**
   - long-term preservation of herbarium specimen images

2. **INFORMATION EXTRACTION**
   - extracting information from images by using Optical Character Recognition (OCR) analysis and further image recognition techniques

3. **KNOWLEDGE BUILDING**
   - exploiting OCR results in a full text analysis tool and offering access and sharing with the whole community
Technical implementation

Museums (producers)
- Community of users
- Community portal
- Images
- Image replication
- Full text indexation
- Access to meta+images
- Monitoring interface

CINES (repository)
- Transfer
- Quality check
- OCR
- Long term archival
- Monitoring

Archives
OCR results
Status and statistics
Service hosting
Collaborations with other EUDAT WP

WP 2
- WP 2.1 Service catalogue
- WP 2.2 Certification
- DSA-WDS implementation
- SLA-OLA definitions

WP 5
- WP 5.3.2 Data Curation and Policies
- WP 5.4.2 Generic Execution Framework
- Embedded HPC instances to execute OCR on-demand

WP 6
- WP 6.2 EUDAT services operations
- WP 6.3 EUDAT project enabling

HERBADROP
Current status

• Achievements
  - service operation and workflows between the B2SAFE submission and the process of OCR data
  - ingestion of more than 3 Millions of images representing 11 M OCR files and 1,78 M of images processed
  - more than 22 TB of volume and 200 000 hours of computation power

• What still needs to be done
Current status

• Achievements

• What still needs to be done
  – finalize the archive function, harmonizing the Metadata model
  – defining a common interface for harvesting specimen catalog including common set of metadata attributes based on Darwin Core standard.
  – improving recognition of hand written text in herbarium images
  – data curation as part of the preservation process (identifying duplicates, or inducing new taxonomic relations, etc.).
  – developing an OAI-PMH endpoint to publish the content the database.
Part two:

2. The reality of being a data partner with HERBADROP
At the RBGE – Plant collection: This is what we are very good at doing.
Plant identification: good at this.
Plant curation: again, good this.
Digitisation: We have learnt in the last 10 years to become good at this.
Online dissemination: An ongoing process, but we have become good at doing this.

http://data.rbge.org.uk/herb
Image size, quality and usability: With 600 dpi TIFF and 7 options of image size we are now good at this.
Image workflows developed at RBGE incorporating:

- image capture
- automated image processing
- metadata recording
- optical character recognition
- quality control
- image streaming online
- archiving

Image workflows: We had to get good at this, it has taken while but we are there now.
Herbarium Data

RBGE project websites
- ADIAC Diatom Image Database
- Apiales Resource Centre
- Dipterocarpaceae Database
- Robert Wight Catalogue
- Sapotaceae Resource Centre
- SE Asian Begonia Database
- Zingiberaceae Resource Centre
- Flora of Nepal
- Webcyta Genetic Flora
- Dryflora
- Cerrado
- Gesneriaceae Resource Centre
- CMEP Plant & Birds Apps.
- Plants & You (Nepal)
- Botanics Stories

RBGE online catalogues
Searchable & downloadable data

RBGE online catalogues

Internally produced resources

External aggregators

GeneBank
BOLD
BioCase
EOL
Global Plants
IDigBio
GBIF
Europeana
CRIA Species Link
NBN
ALS

External end users

Biodiversity scientists
Taxonomists, habitat restoration, species checklists, geneticists etc.

Providing data aggregators: Once again had to get good at this, if we wanted people to use our data.
This is where we are less strong – coding and command line interaction.
Preserving this – a high resolution image of a herbarium sheet is great, but we want more.
Georeference the collection site, database the collector and collection date; as here is physical proof that this material was at a site on a particular date.
Dets: (determinations) taxonomic opinions on the sheet, all with different levels of validity – OCR is never going to the complete answer for our material.
As a community we use unique identifiers for our material, we need these identifiers to connect material.
We need to have stable identifiers to link out to other sources
We need to have stable identifiers to link out to other sources.
We need these other sources to be funded and stable.
Robert Cubey
Plant Records Officer
Royal Botanic Garden Edinburgh

r.cubey@rbge.org.uk
@RBGE_plant_rec
+44 (0) 131 248 2948