digitization of P Herbarium

Switching to the fast track: Rapid digitization of the world's largest herbarium

Botany 2011 – Colombus, Ohio

Marc Pignal, Henri Michiels

11th of July, 2012

The French Museum, an old institution

- Founded in 1635 (at that time the Royal garden of medicinal plants)
- 70 million collection specimens
- 15 locations, 2000 people
- 350 researchers
- 400 students (Master and PhD)

digitization of P Herbarium

Renovating the Paris Herbarium

An opportunity to digitize the entire collection

The Digitization Project

- The original objective was to renovate the building, in order to raise its capacity from 6 to 10 million specimen sheets
- To do this, we had to move away the entire collection during the works
- It was an opportunity to digitize all the sheets of the Phanerogams and ferns

  (ca. 7 million specimens)

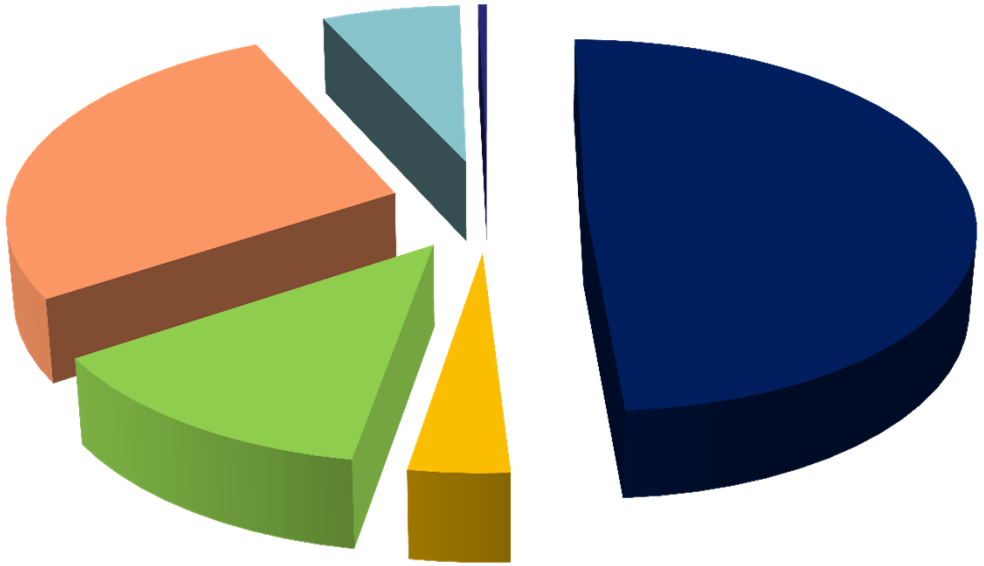
Budget

Overall project cost: 24,5 Million €

(ca. 30 Million USD)

- Building renovation
- Movers
- Mounting specimens (partial)
- Reconditioning, digitization and sorting
- Supplies
- Storage

6 700 000 €
8.5 million USD



2D Digitization is cheap

- the cost of digitization is marginal compared to the full project
- full specimen processing (moving, sorting, reconditioning, new furniture) $1,5

→ digitization and name processing $0,1

The renovation cycle

Herbarium

1

5

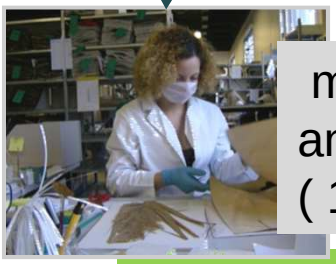
Digitization
Reconditioning
ordering and
packing

Sorting the backlog

2

3

Industrial partner (Océ)



mounting and prefiling (1 M sheets)

private company

Storage (moving company)

4

Rapid digitization of P Herbarium

Before ....

… And after

Rapid digitization of P Herbarium

digitization of P Herbarium

The workflow

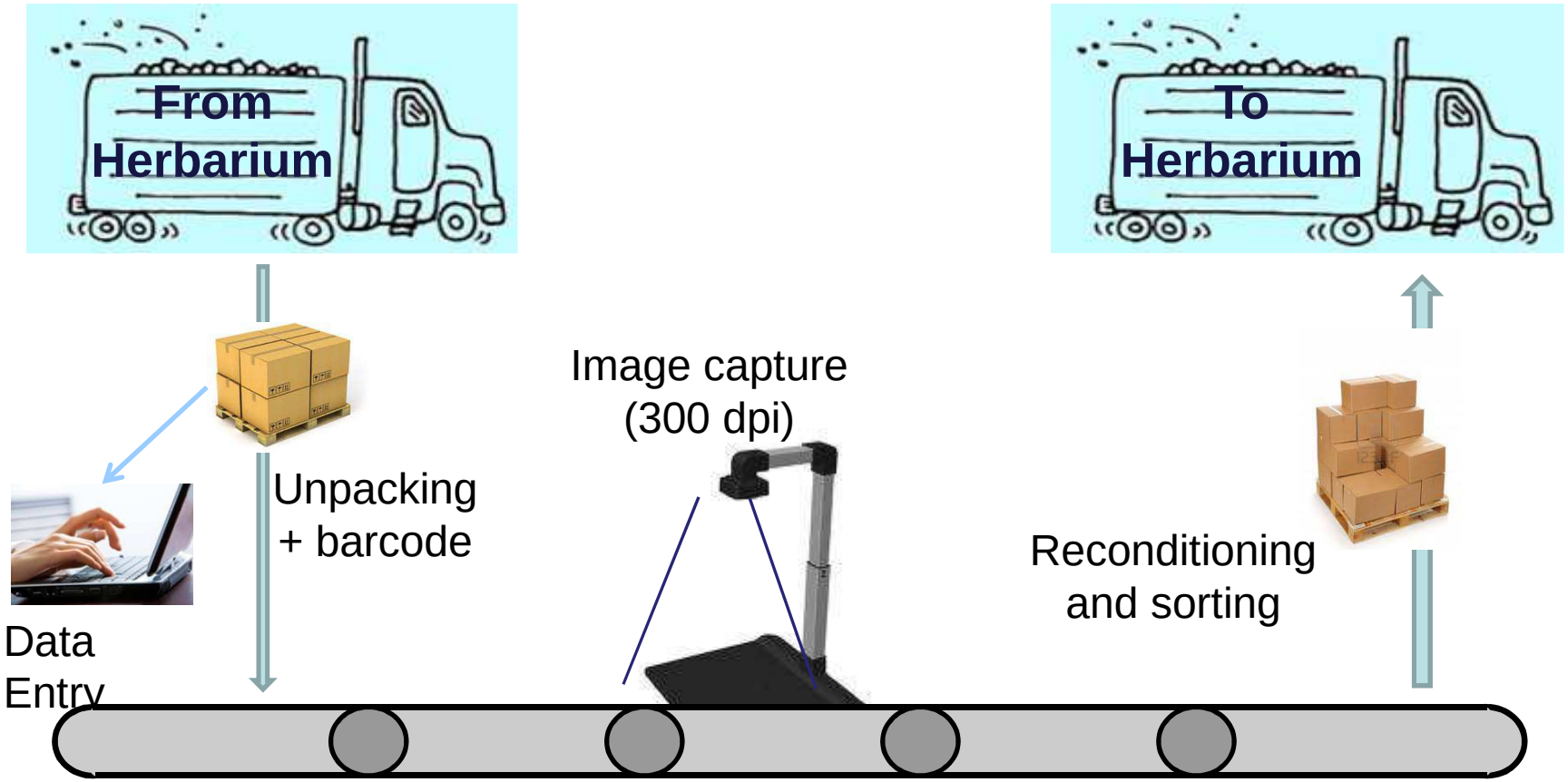
Digitizing, reconditioning and sorting

An industrial process

- We selected a private partner who had to set-up a dedicated workshop
- 20 people working in two shifts
- Planned production rate: 17 000 sheets per day over 24 months

The digitization site

Workflow overview



1 – Delivery (1)

The moving company carries the specimens to the facility where they arrive in clearly labeled boxes

1- Delivery (2)

- 1600 boxes delivered monthly, part of them includes the unsorted backlog
- Boxes receive a tracking barcode

1- Delivery (3)

1- Delivery (4)

- The Museum provides a “taxonomic file” describing the content of the boxes
  - box number
  - family name and APG number
  - genus name and serial number within family
  - geographic area

1- Delivery (5)

- This information is inserted in the contractor's Information System and used along the industrial process (labeling, sorting, quality assurance)

2- Folder processing (1)

For each genus folder, the operator:

1. replaces the old jacket with a new one (color according to region)
2. types the first letters of the species name and selects the name from the taxonomic list (family, genus, species, authors, ID=taxon number)
3. prints a label with barcode and identification information, and sticks it on the folder

2- Folder processing (2)

3 – Specimen Digitization (1)

- Datamatrix and barcode are stuck on each sheet
  - Datamatrix: for tracking purposes
  - Barcode: specific to Museum and to international herbarium standard
- The specimens are placed three by three on a tray

3 – Specimen Digitization (2)

3 - Specimen Digitization (3)

- The tray is placed on a conveyor belt
- The tray is scanned
- The scan is checked (framing and focus)
- At the end of the chain, the barcode is read to check if all specimens are back in the folder

3 - Sp (4)

The Digitization Bench

4 – Reconditioning (1)

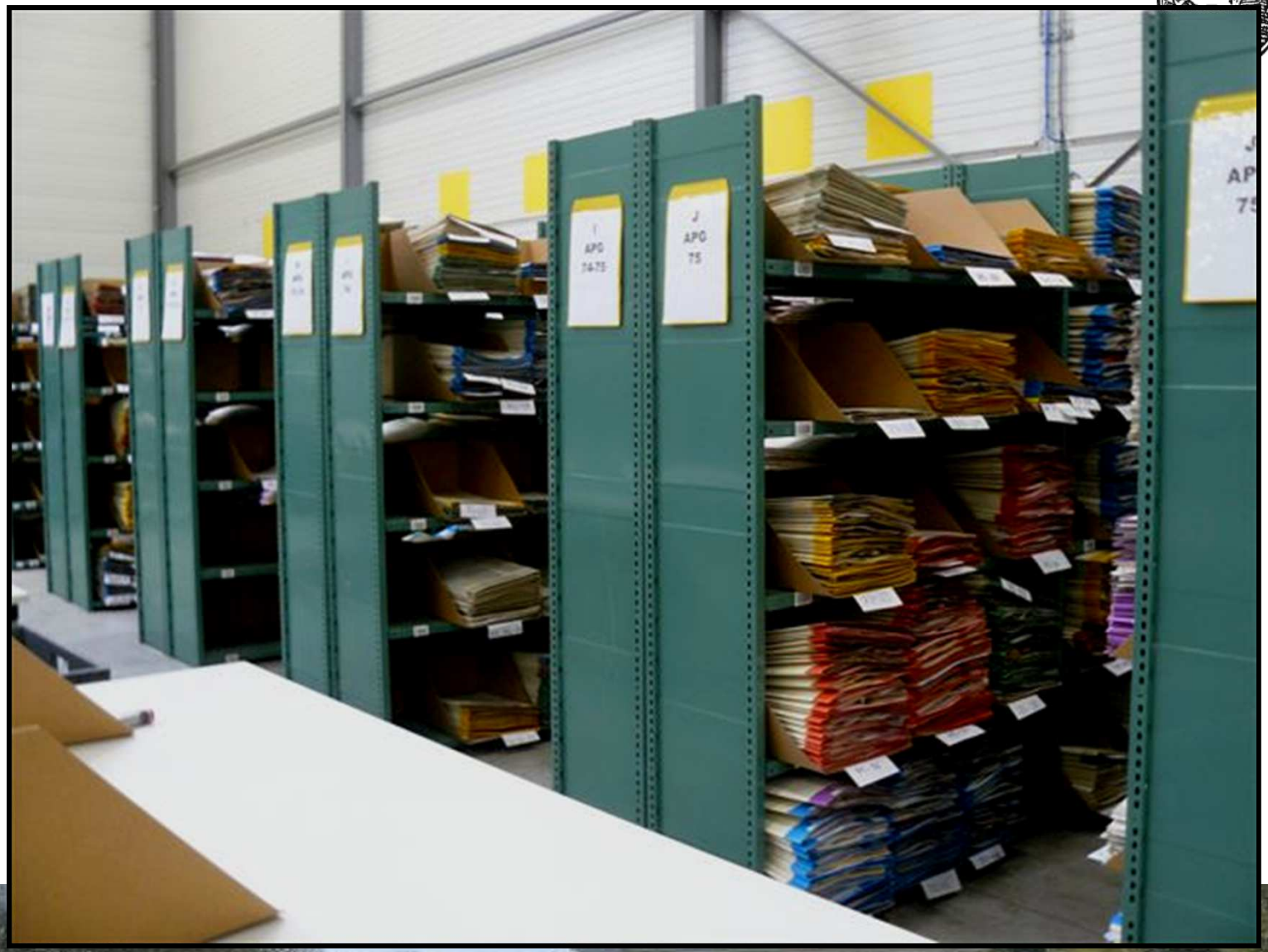
- After scanning, each sheet is placed in an individual paper protector
- The barcode of each specimen is read, allowing the system to check if all specimens are back in the right folder
- The folders are stored in a box before sorting

4- Reconditioning (2)

5- Sorting 1 (by genus)

- This sorting consists in storing specimens by family and genus names
- The operator puts the genus folders in boxes and places them on shelves according to the family and genus numbers (the shelves are labelled in advance by the contractor)

5- Sorting 1 (by genus)

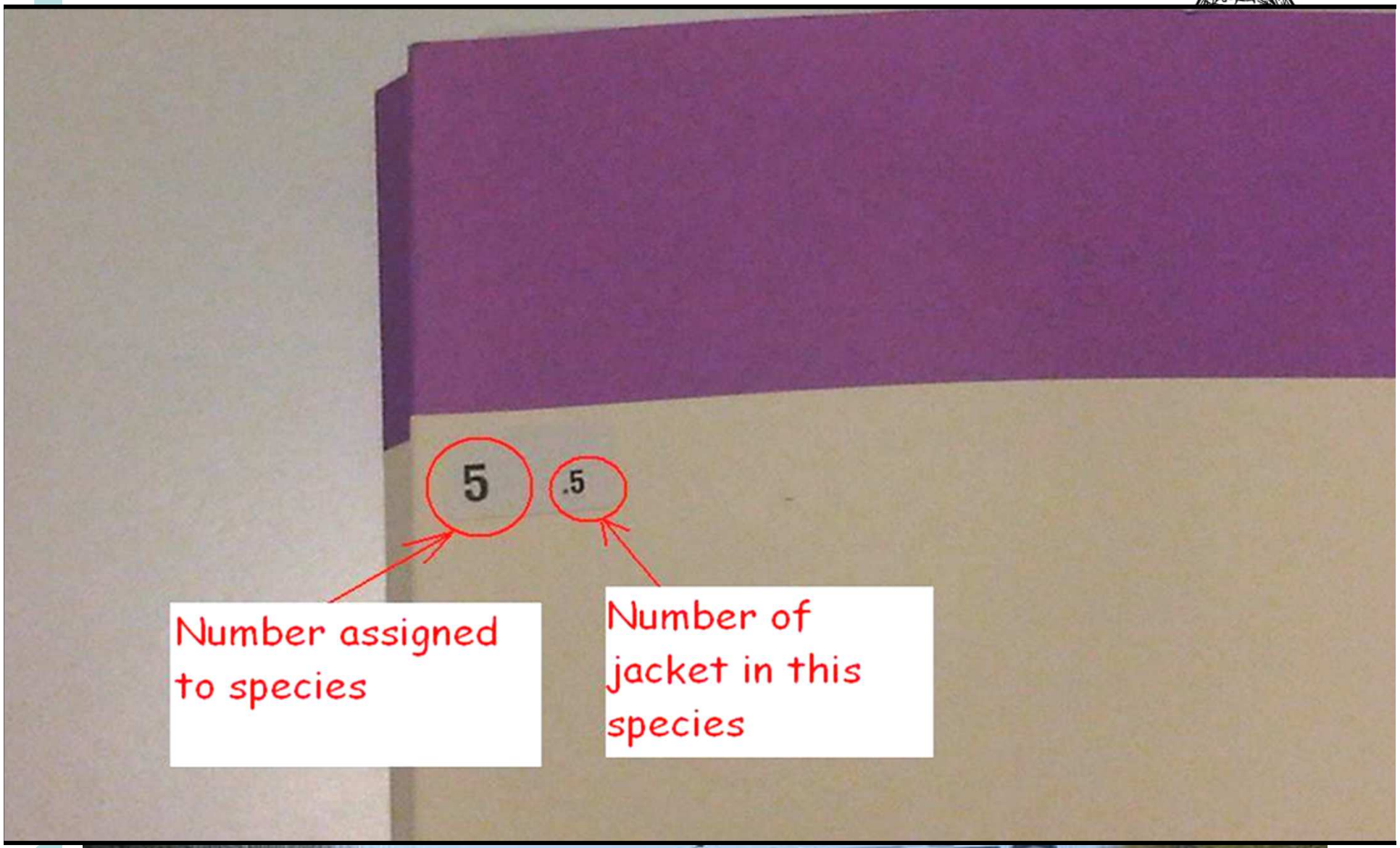


6- Sorting 2 (by species)

- Inside each (genus, geographic sector), the operator sorts the folders by species

6 - Sorting 2 (by species)

- The operator reads the barcode on each folder
- The system displays the species name and assigns a number which is printed on a label
- The label is sticked on the folder, which is then stored on the shelf with the same number


5
.5
Number assigned to species
Number of jacket in this species

7 – Packing, transport and final storage

- The folders are packed in boxes, carried back to the Herbarium…

… and finally installed in the new compactors

7 – Packing, transport and final storage

digitization of P Herbarium

Scanning Resolution and Image Format

Production of images

- The conveyor belt passes the specimens under a **bidirectional scanner** which produces 11x17” (**A3**), **300 dpi** images
- TIFF files are saved offline (one production day per disk of 1 TB)
- JPEG’s are made for online use
- Filenames are generated from the barcode number read on the images

Image size

- One TIFF image weighs 50 MB
- One JPEG is 5 MB. This compression rate was chosen to have the same level of details as with TIFF (only colour is slightly changed)
- **This choice is a technico-economic trade-off**
- For 10 million images:
  - TIFF files represent 500 TB
  - JPEG files represent 50 TB
  - Database information represent <100 GB

Handling TIFF data

- We cannot afford « live » storage of 500 TB
  - … and even 1 Po with redundancy ! $$$
  - With a lot of energy consumption and heat dissipation for rarely accessed images
- We will start using tape storage soon, with HSM* software

  * hierarchical storage management
- For the time being, USB disks are stored in the Museum's library

Exception for Types

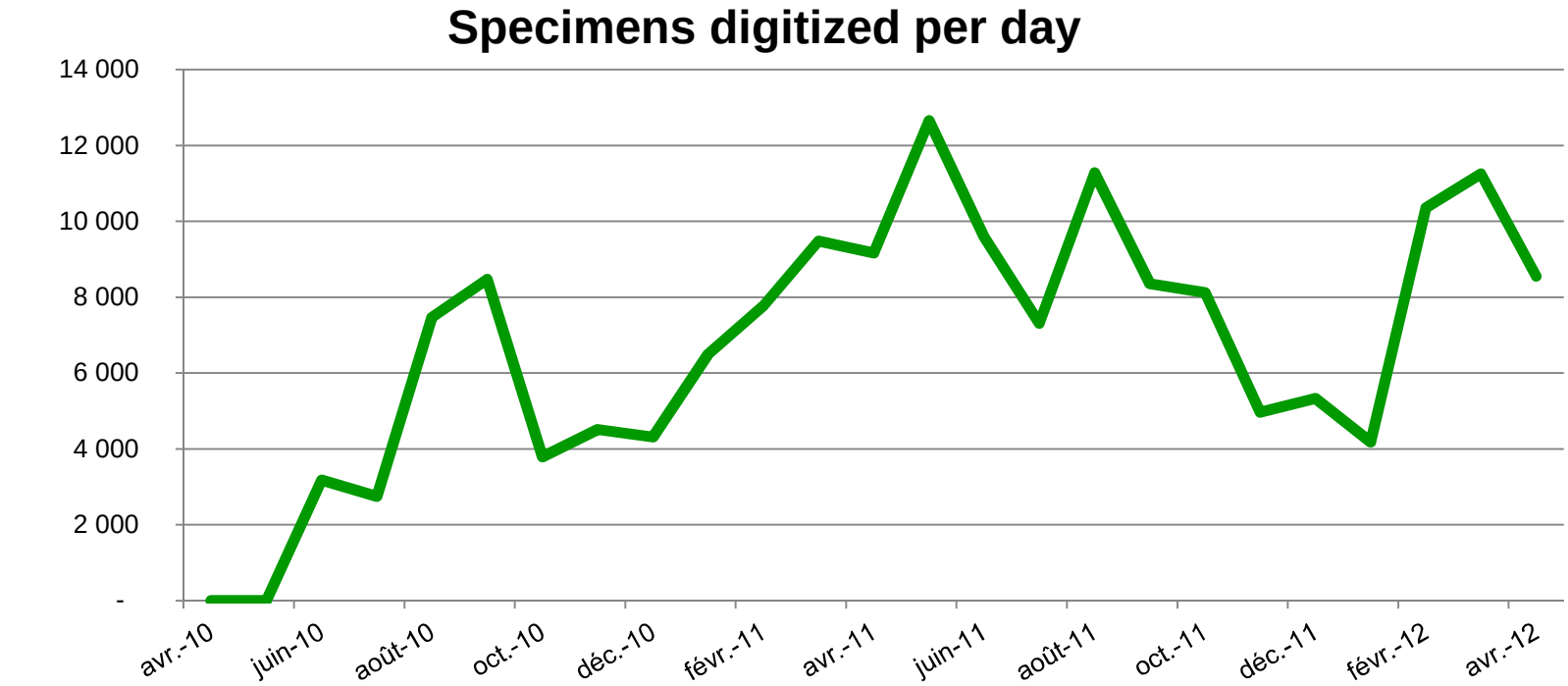
- The types and historical collections are not part of this industrial process
- They are manually digitized on-premises at 600 dpi (200 MB in compressed TIFF)
- This process was initiated by the Mellon foundation in 2004
- We now have 300 000+ type images

What we've achieved and learned …

Achievements

- 4 million specimens processed between June 2010 and June 2012
- Images and data are of good quality

Fast but ... not fast enough



Software and quality assurance

- A continuous control is mandatory
- There is more software needed for ensuring traceability and detecting failures than for data acquisition.
- Fast web publication of images allows a broader audience to perform quality control.

How to ensure quality in mass digitization?

60 000 images produced each week

1

1% of the production checked (ca. 600 images)

2

Samples are distributed among botanical staff

3

Checking:

- Focus
- Data quality
- Barcode number
- Barcode location

4

Working with a contractor

- Culture clash

ROI	⟷	quality
speed	⟷	exhaustivity
robustness	⟷	specifity

- **Quality control is a key point** to make sure that scientific excellence governs the industrial throughput

What's next ?

Digitization is just a very first step…

Virtual herbarium

- Extract text information from pictures to database (crowd sourcing, OCR, …)
- Transpose working methods from physical to on-line to search, compare, determine, …
- Allow “virtual visits” with working tools adapted to this new interface
- …

Which means many projects to come !

Thank you !

- Direction des Collections / Systematics Dpt

  Marc Pignal - pignal (at) mnhn (.) fr

- DSI (Information Systems)

  Henri Michiels - michiels (at) mnhn (.) fr

Muséum national d'Histoire naturelle, Paris (France)