

# SwissCollNet

## Data aggregator

Digitization processes in natural history collections:  
from drawers to data publication.

**Nils Arrigo**

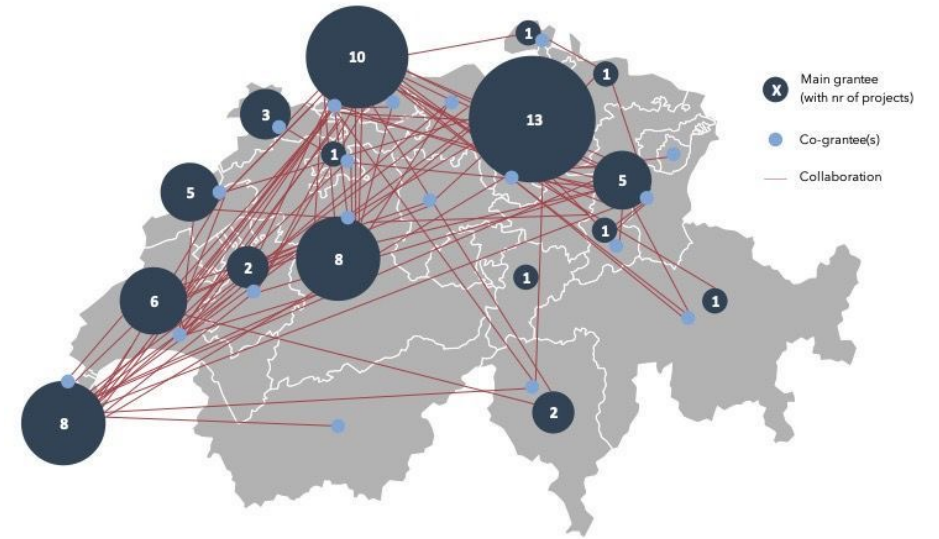
Info fauna  
2024.01.19



The screenshot shows the website for the SwissCollNet network. At the top left is the logo for 'scnat académie des sciences naturelles'. To the right are language options 'de fr it en' and navigation icons. The main header is a red banner with the title 'Réseau suisse des collections d'histoire naturelle (SwissCollNet)' and a short description in French. Below the banner is a navigation menu with buttons for 'Implementation', 'Running projects' (highlighted in red), 'Collection data', 'Exchange', and 'Organisation'. A breadcrumb trail reads '> Page d'accueil > Running projects'. To the right of the breadcrumb are social media icons for Facebook, LinkedIn, Twitter, and a share icon. The main heading is 'Running projects', followed by the sub-heading 'The SwissCollNet projects are launched!'. The text below states that 68 projects are financially supported, with 41 based on cooperation between collection institutions. To the right, a section titled 'Distribution of institutions, disciplines and funds of SwissCollNet projects' mentions 115 main- and co-grantees from 53 institutions. Below this is a map of Switzerland with a network diagram of nodes and connections. A legend indicates that dark blue nodes represent 'Main-grantee (with or without)', light blue nodes represent 'Co-grantee(s)', and red lines represent 'Collaboration'. The caption below the map reads 'Image : SwissCollNet'.

# Missions of the data aggregator

- **Harmonize the flow** of natural history collection data among Swiss and international institutions
- **Enhance the visibility and accessibility** of specimens, collections and institutions, by publishing the data on national and international portals



# Challenges and solutions

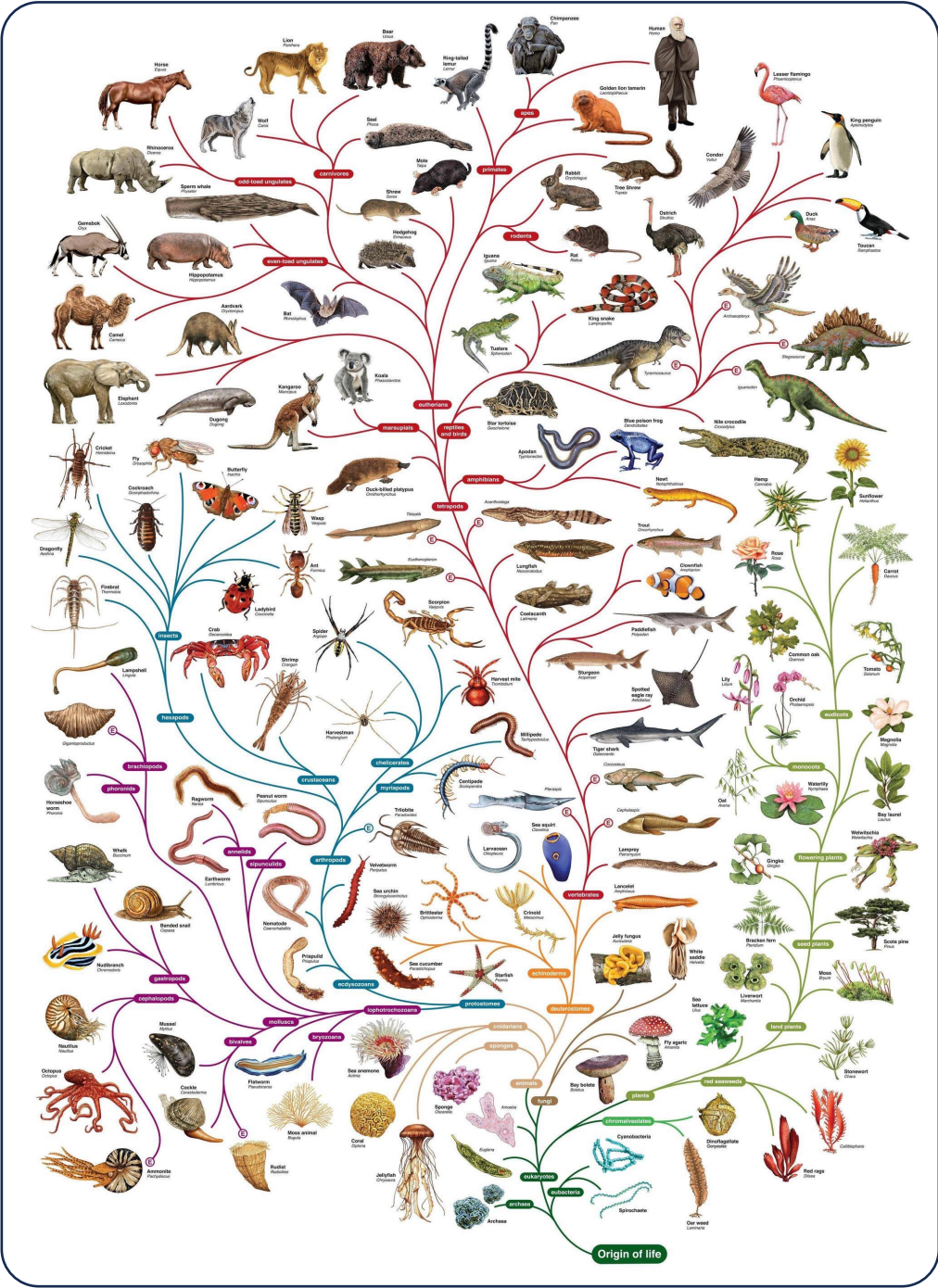




# Challenges

The information to gather...

- ... covers many types of objects and concepts ...





# Challenges

The information to gather...

- ... covers many types of objects and concepts ...
- ... uses heterogeneous and ambiguous representations ...



« Legatee »

« Collector »

« Donator »

« recordedBy »



«1 record = n specimens»



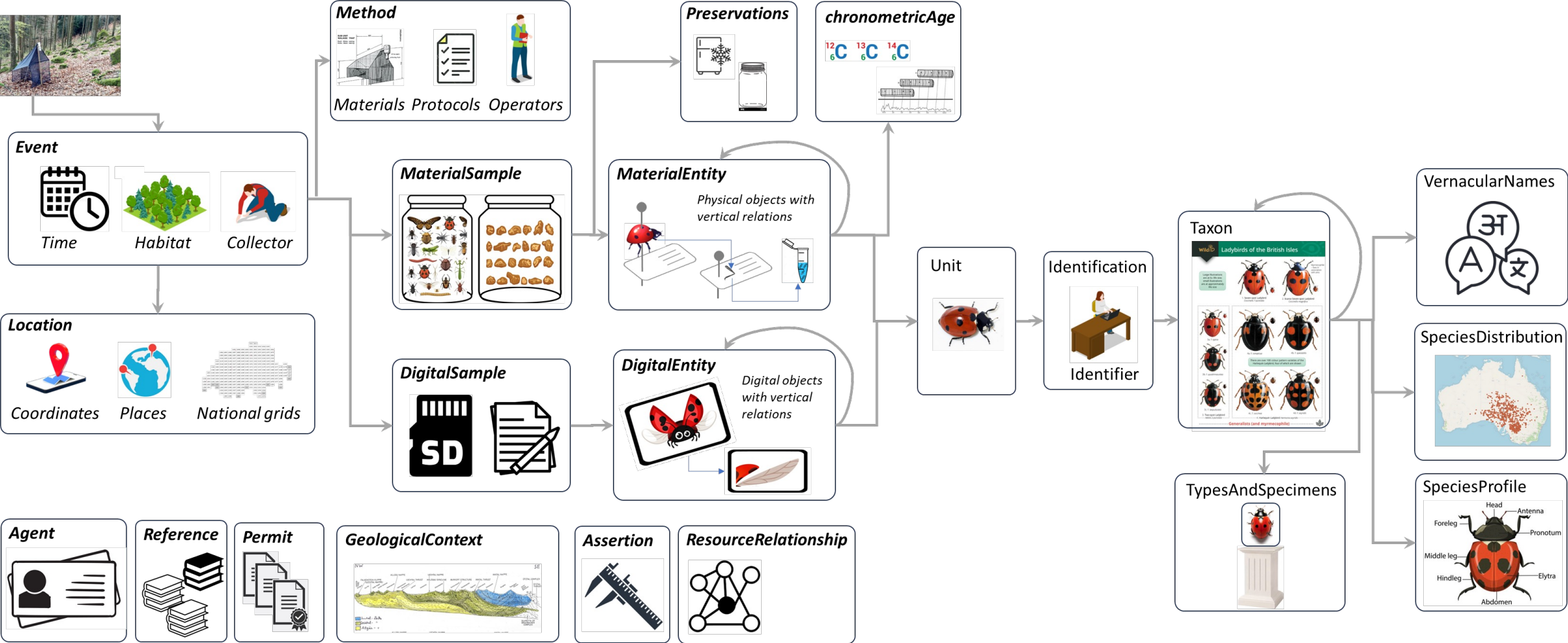
«1 record = 1 specimen»



*Elaphe taeniura* (Cope, 1861)

*Orthriophis taeniurus* (Cope, 1861)

# Solution 1. We use a general data model\*

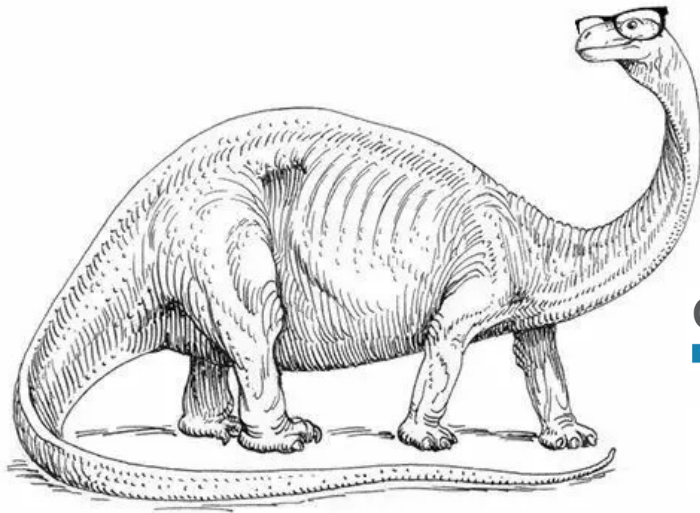


\*Includes all GBIF extensions & borrows concepts to ABCD

# Solution 2. We rely on publicly available controlled vocabularies

What do you call a dinosaur  
with an extensive vocabulary?

**A thesaurus.**



PubMed®

GBIF

Catalogue of Life

swisstopo  
+ + +

Grammarly Quotes

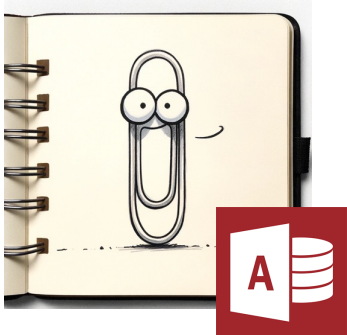
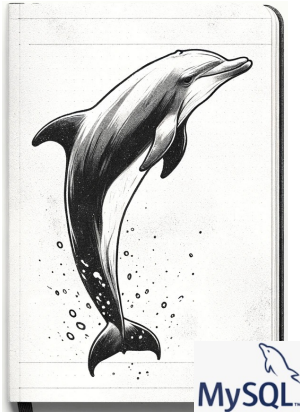
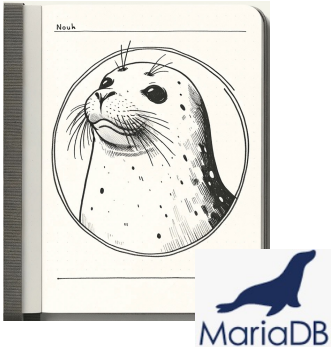
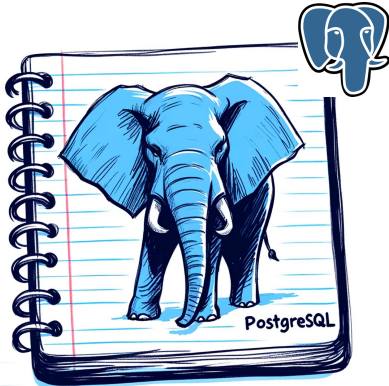
- Propelled by live data sources
- Maintained by a wide scientific community
- Available via automatic means (APIs)
- Available within SVNHC as an ENCODING service (manual checks remain needed)
- Scalable (we can add catalogs as per our needs)



# Challenges

The information to gather...

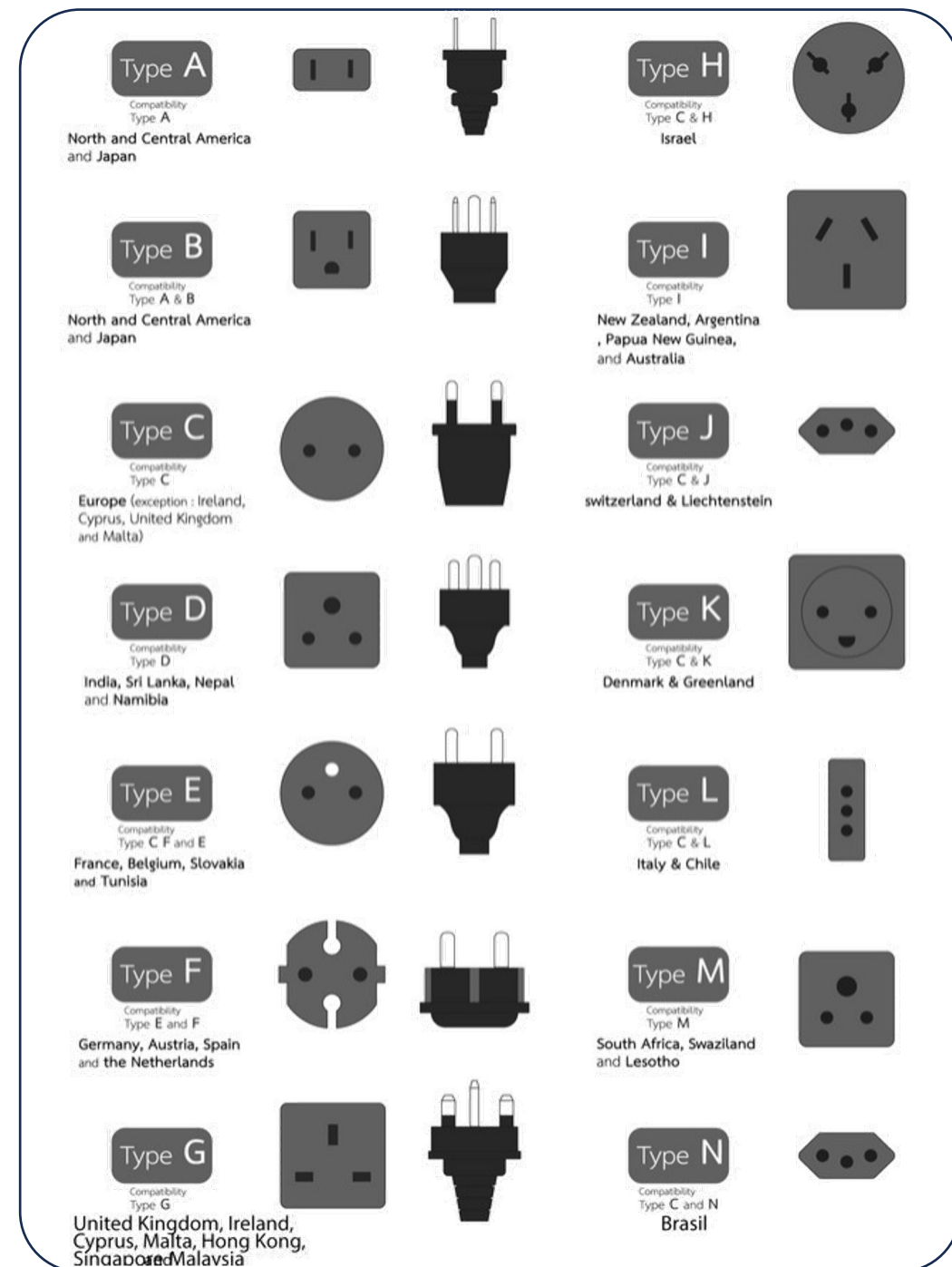
- ... covers many types of objects and concepts ...
- ... uses heterogeneous and ambiguous representations ...
- ... is stored in heterogeneous database systems ...



# Challenges

The information to gather...

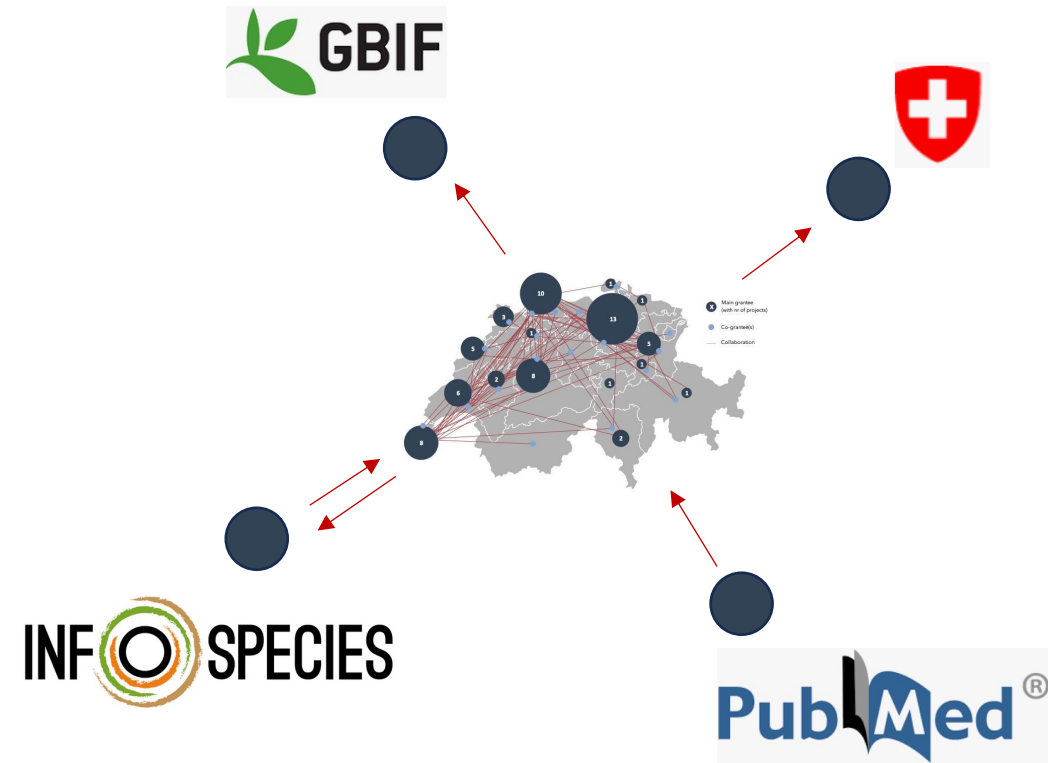
- ... covers many types of objects and concepts ...
- ... uses heterogeneous representations ...
- ... is stored in heterogeneous database systems ...
- ... hides behind access barriers across 57 institutions ...



# Challenges

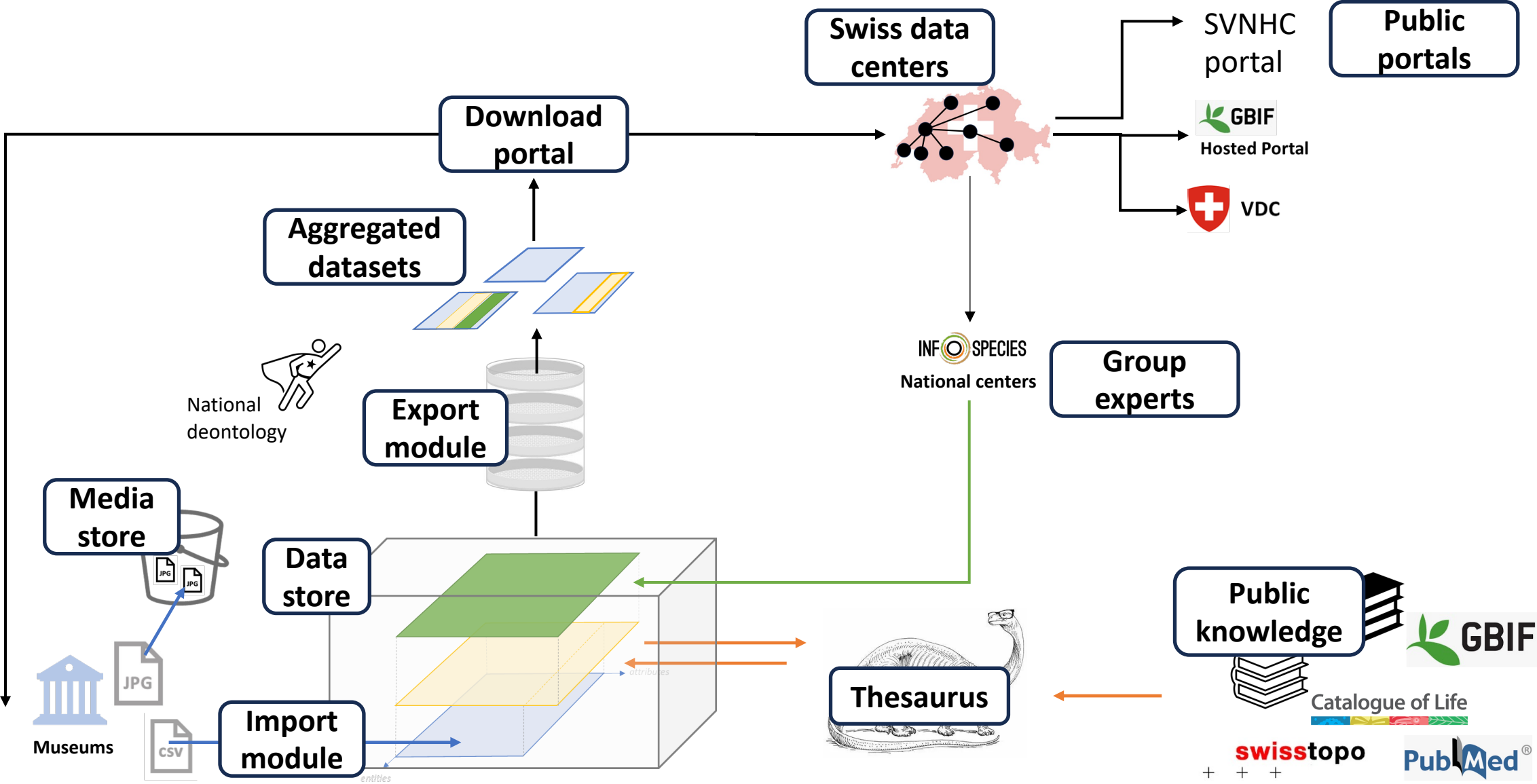
The information to gather...

- ... covers many types of objects and concepts ...
- ... uses heterogeneous representations ...
- ... is stored in heterogeneous database systems ...
- ... hides behind access barriers across 57 institutions ...
- ... flows throughout a wide network of actors ...





# Solution 3. We host the data in a layered + interconnected system

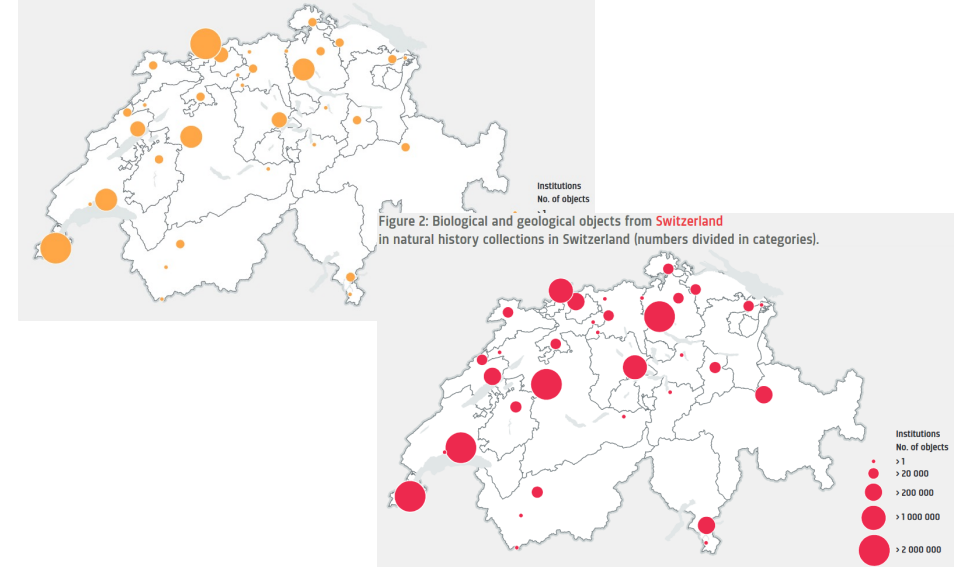


# Challenges

The information to gather...

- ... covers many types of objects and concepts ...
- ... uses heterogeneous representations ...
- ... is stored in heterogeneous database systems ...
- ... hides behind access barriers across 57 institutions ...
- ... flows throughout a wide network of actors ...
- ... and represents large volumes

Figure 1: Biological and geological objects of national and international origin in natural history collections in Switzerland (numbers divided in categories).



**60.6 million objects**

of which

**10.7 million digitalized**

completed by

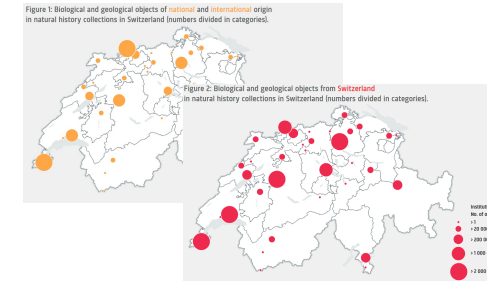
**2.5 million digitalized  
via SwissCollNet**

# Challenges

The information to gather...

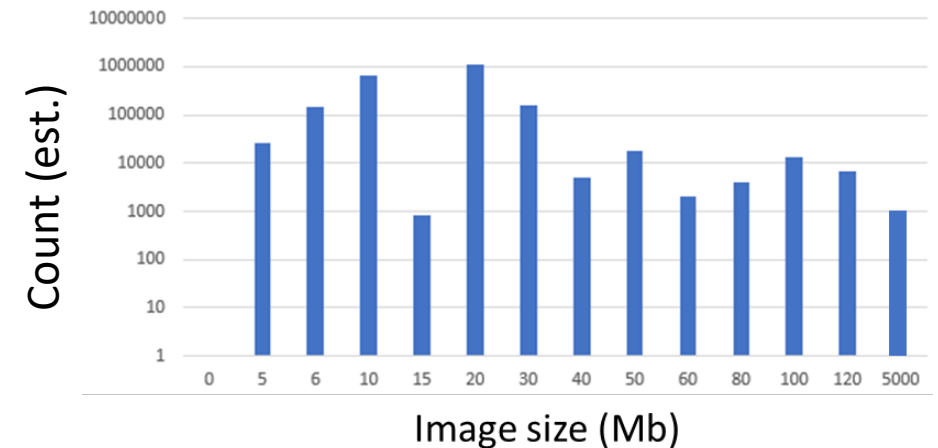
- ... covers many types of objects and concepts ...
- ... uses heterogeneous representations ...
- ... is stored in heterogeneous database systems ...
- ... hides behind access barriers across 57 institutions ...
- ... flows throughout a wide network of actors ...
- ... and represents large volumes

\*for original images, established as per Media survey of December 2023, hosting fees account for one S3 Store + one backup



60.6 million objects  
of which  
10.7 million digitalized  
completed by  
2.5 million digitalized  
via SwissCollNet

Worth 45Tb of data\*



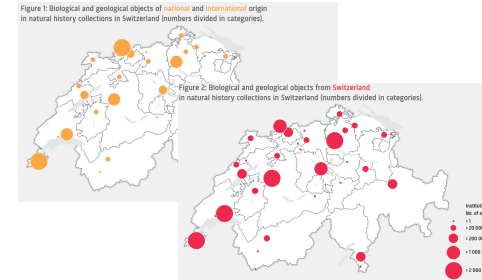


# Challenges

The information to gather...

- ... covers many types of objects and concepts ...
- ... uses heterogeneous representations ...
- ... is stored in heterogeneous database systems ...
- ... hides behind access barriers across 57 institutions ...
- ... flows throughout a wide network of actors ...
- ... and represents large volumes


\*for original images, established as per Media survey of December 2023, hosting fees account for one S3 Store + one backup



60.6 million objects  
of which  
10.7 million digitalized  
completed by  
2.5 million digitalized  
via SwissCollNet

Worth 45Tb of data\*

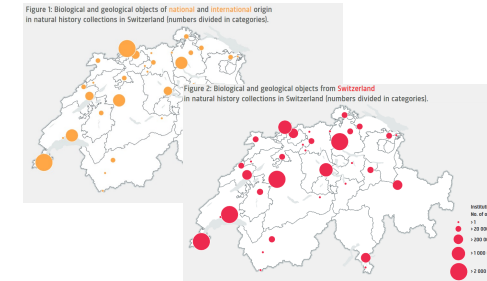


	11'615 CHF / yr
	11'250 CHF / yr
	14'400 CHF / yr
	14'940 CHF / yr
	16'830 CHF / yr
	17'280 CHF / yr

# Challenges

The information to gather...

- ... covers many types of objects and concepts ...
- ... uses heterogeneous representations ...
- ... is stored in heterogeneous database systems ...
- ... hides behind access barriers across 57 institutions ...
- ... flows throughout a wide network of actors ...
- ... and represents large volumes



60.6 million objects

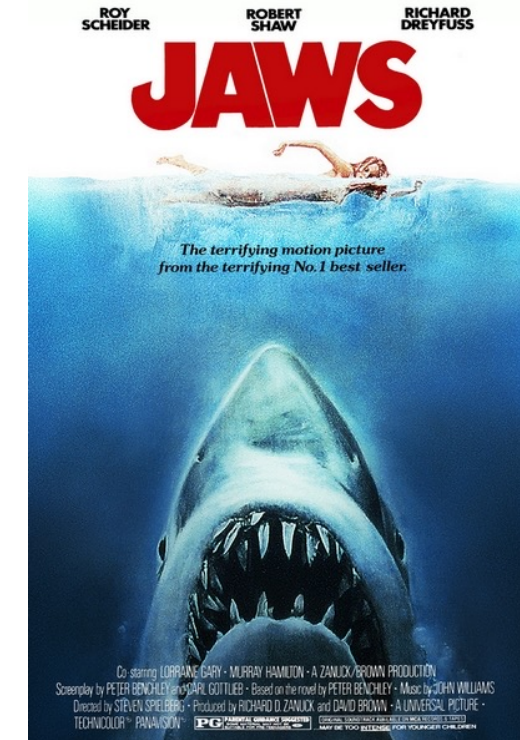
of which

10.7 million digitalized

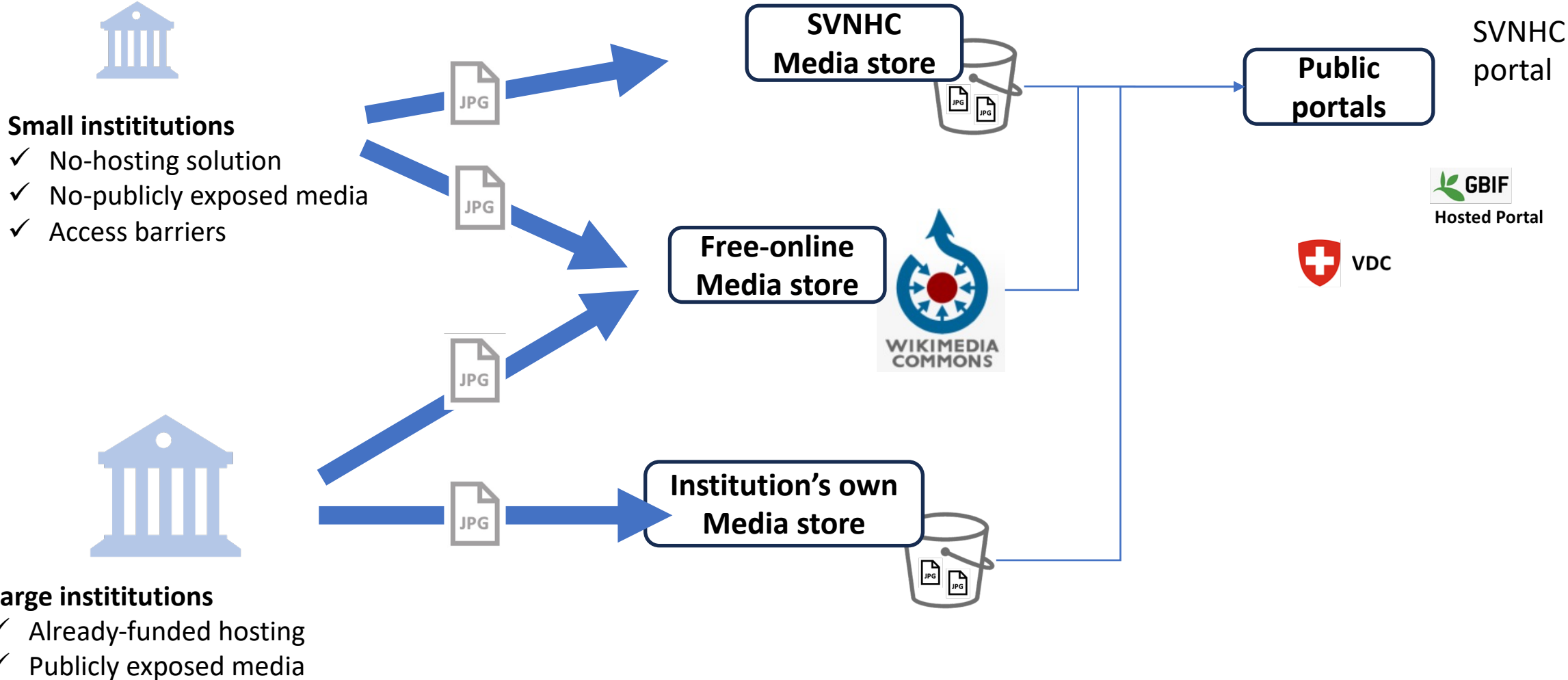
completed by

2.5 million digitalized  
via SwissCollNet

How much data is hiding here ?



# Solution 4b. Spread the burden of hosting media files



# Solution 4b. and/or accept lower resolutions

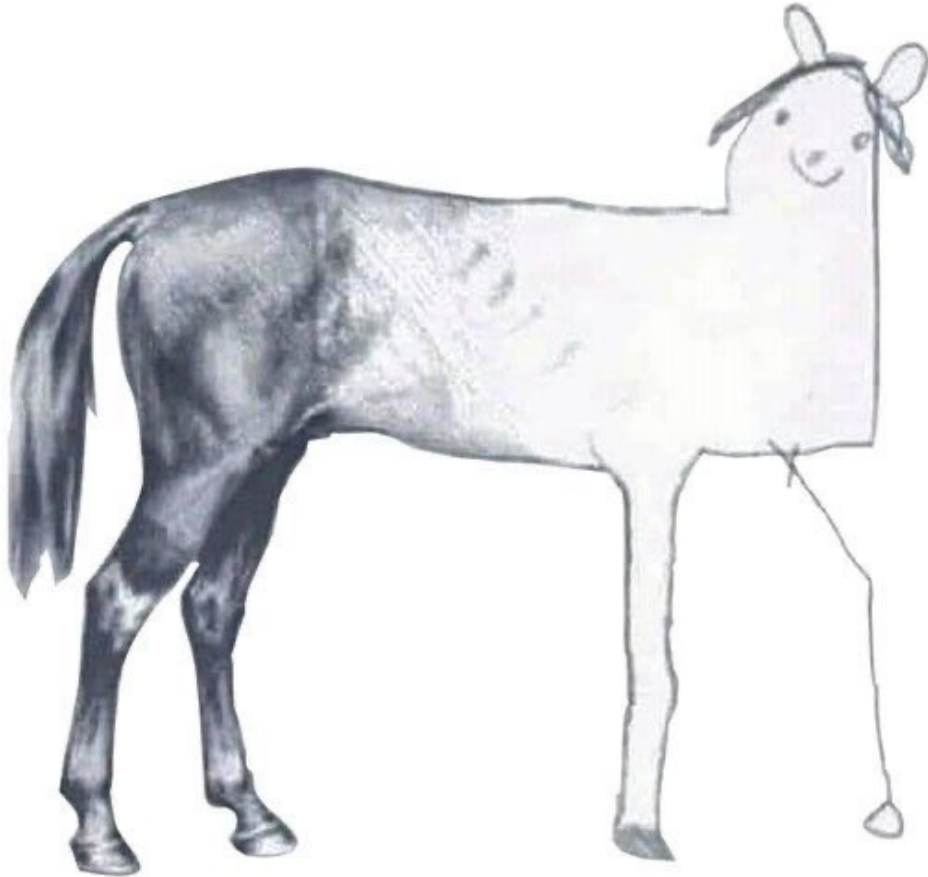


Drawer scans : an example at 6Mb / image  
(original : 15Mb / image)

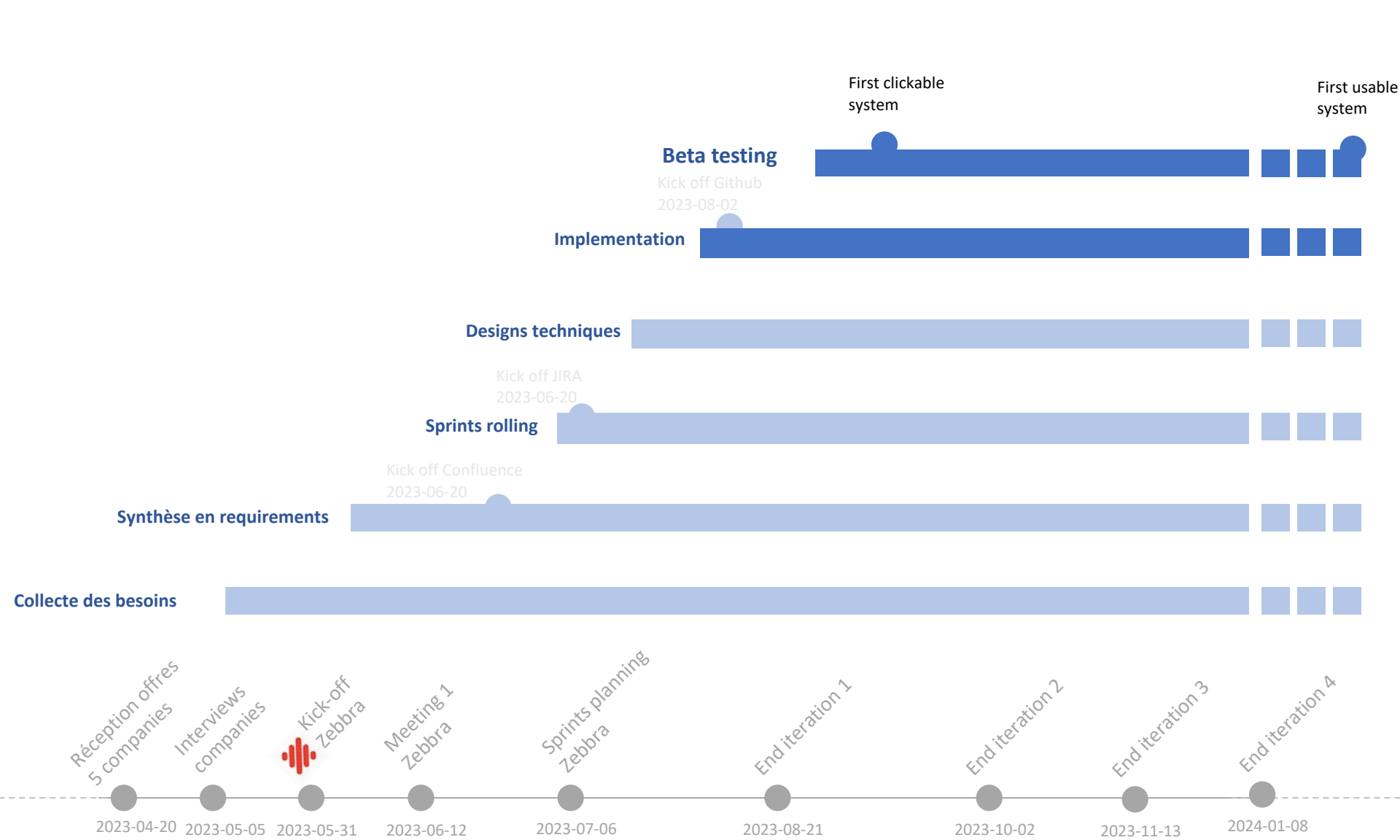




# Timeline



# Timeline



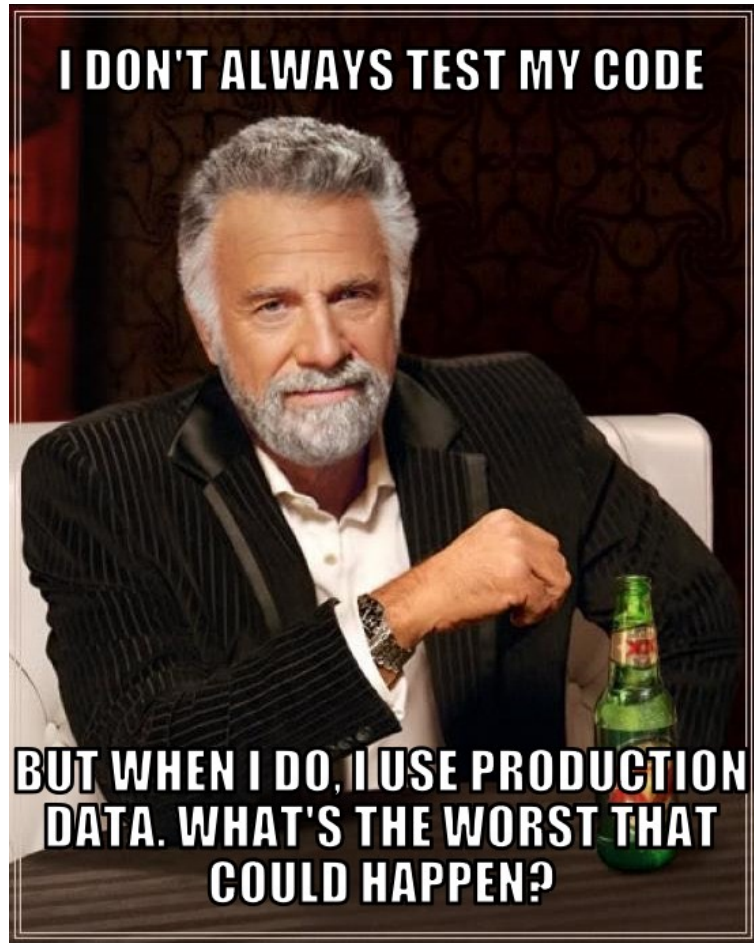
imgflip.com

JAKE-CLARK.TUMBLR

# What to expect now ?



# Digitalization teams with finished projects



Please reach out to us with demo files

We want to see representative datasets !



# Testing team: we are getting there !

The screenshot shows a web dashboard with a dark theme. On the left is a sidebar with navigation links: Dashboard, Collections (highlighted with a mouse cursor), Records, Imports, External Resources (Github, GBIF, scnat), and Internal Resources (Dashboard, GraphQL Playground, HexDocs, Storybook, Rdoc JSON API). The main content area is titled 'Dashboard' and contains eight data cards in a 2x4 grid:

Amount of Collections <b>0</b>	Total Records <b>0</b>	Digitization Progress <b>74%</b>	Records Published <b>3072</b>
Records Reviewed <b>1207</b>	Last Contribution <b>13.11.2013</b>	Open Reviews <b>27</b>	Contributors <b>87</b>

At the bottom left, the browser address bar shows 'localhost:4000/collections'. The top right corner has a language selector 'EN' and a desktop icon.

**Thank you for your time and attention!**