Technical discussion

Subject :

- **taxonomy**
- **new data types**
- **index new field**
- **metadata publishing**

Sweden : taxonomy - want to use more than 1 taxonomy - launch the indexation against several taxonomies

6 institutions will use their national checklist
7 institutions will need to accommodate their taxonomy list

taxon works - tools - ruby on rails - webservice - Matthiew and Tim have presented at the previous TDWG

Tools to manage Taxonomy (other than ALA defaut) : none ?

A use case in Sweden is to merge the national official taxonomy with a bacterial taxonomy
Taxonomic list include on national one ?
Not included (Sweden)

molecular data =/= occurrence data = > expand dwc
-> on higher level than the ALA tools ; more generic issue

BGBM (?) Berlin incorporate different taxon concepts -> how to prepare material samples

France : connect data outside DwC and indexing new field

In France : issue with data with other fields who aren't in the DarwinCore standard

5 institutions have sensitive data/species

Portugal will need to have other ???

Biocache has the abilities to generalize coordinate

4 institutions would like it

Norway : the government manage sensitive information on another system

CAS : several users - several roles

UK : sensitive data - don't display or interact with sensitive data - if they have the right, they can download more information - depending of species but also location
Possibility to document the UK process ? it's a fork of exitant ALA component

Brazil : sensitive access - different taxonomy

Guatemala : difficulty to get the attribution correct => pb mapping collection / institution / data resource

Jeremy : python script for insert information in MySQL DB
Share existing Python Script with the community (if any) or develop new one
People are interested in using Python script

Share all script with the community : they are at least one person who are interested in it
Talk about each national development : people can help you or use what you develop and improve it.

oai-pmh protocol : https://www.gbif.org/developer/registry#oai-pmh : metadata publishing

Norway : (@UK) SIO has an ontology terms for life status
http://semanticscience.org/resource/SIO_010057
SIO:010057 could be suggested added to an application schema for the IPT (and thus Darwin Core archives...)
https://www.gbif.org/grscicoll
DCAT service from the IPT: https://github.com/gbif/ipt/pull/1185#issue-41317353 eg: https://data.gbif.no/ipt/dcat
See also the OAI-PMH service from the GBIF registry (only EML and Dublin Core, not DCAT?): https://www.gbif.org/developer/registry#oai-pmh

Portugal : enough resource for infrastructure but not people.
National GBIF node : part of the national research infrastructures (RI) roadmap through PORBIOTA. Don't need to install server. Uses cloud service provided by another national RI INCD. INCD links at Iberian level with IBERGRID, and at european level with European Grid Initiative (EGI). Ideally, infrastructure based in Portugal mirrored in Spain. New server to be acquired will be hosted by INCD infrastructure. Start implementing in other hub based on geographic instead of politics regions - help other countries in which we have a relationship. Deal with sensitive information with sensitive political issues, but agreements can be made.

------------------------------------Begin Infrastructure description--------------------------------------

Sweden: https://github.com/bioatlas/ala-docker can be deployed in various Cloud infrastructures such as Google Cloud Platform, Amazon, Microsoft Azure and Open Stack using Docker Swarm and Docker Machine. Currently https://beta.bioatlas.se is deployed using SUNET Cloud - Swedish University Network Cloud through Open Stack delivered by https://www.safespring.com/tjanster/safespring-compute/ and we are migrating the https://bioatlas.se to use the same IaaS provider.

France (GBIF France) - Benin - Togo : Share back-end but three different front (portal interface - data-hub) - Togo is on the process of configuration
6 VMs who hosted the ALA component (Spatial, Biocache, Collectory, Hub, Cassandra, SOLR) on the French node of the EGI cloud network

France (UMS PatriNat) : one VM for all component. On a server located on the French Museum IT local ; on pre-Alpha now

Canada - AAFC - Private Cloud with OpenShift or Public Cloud (TBD, ALA not yet implemented). Getting them is quite a complicated process.

USA - Chicago - Field Museum: Ubuntu Server LTS in VMs in cloud service (TBD, ALA not yet implemented). We have enough compute resources and money to handle the implementation. We shouldn't have any problems doing an implementation.

USA - Vermont / Vermont Center for Ecostudies:
Today: 2 Ubuntu-16 AWS EC2 VMs: estimated annual cost $7K
IPT (t2.micro 30G SSD)
ala-demo (core) (t2.2xlarge 300G SSD)
Tomorrow: Perhaps use docker containers to spread components across smaller (including more free micro) VMs to save hosting costs.

UK: ~45 Ubuntu servers on AWS EC2, and ~5 on Azure. About 10 of these are fairly large (Cassandra x4, SOLR x4, BIE Index, Layers), the rest are small front-end component servers. Total diskspace ~11TB. We are investigating options to rehost the Cassandra and possibly SOLR servers to bring down costs. (currently ~GBP5k/mnth)

Taiwan: Biodiversity Research Center, Academia Sinica. 6 CentOS servers on premises and utilising AWS EC2. (ALA not yet implemented)

Argentina: All our servers are hosted at our data center (Mincyt)

- 3 VM on production (Argentina infrastructure):
    - **ALA modules (biocache, collectory and image) and solr**: Ubuntu 14, 32 GB RAM, 100 GB HD.
    - **Cassandra:** Centos, 12 GB RAM, 100 GB HD.
    - **Image**: Ubuntu 14, 4 GB RAM, 100 GB HD.
- 2 VM on development with spatial and species module.

Spain: ~14 Ubuntu servers (Openstack)

- Spanish Atlas infrastructure schema
- Servers are hosted at IFCA-CSIC, which is part of the GRID initiative. We get this service for free as they get money from EU to provide supercomputing facilities.

- We don’t have control of the barebone servers infrastructure just perms to manage our virtual servers, so we have had lots of service problems because the IFCA infrastructure is not very stable. For instance, since yesterday the openstack login does not work (so we cannot give you exact resources usage stats); this year we had our biocache main server without network for more than a week, etc.

INBio - **Costa Rica**: 5 Ubuntu VMs in cloud provider (https://www.serverpronto.com/, planning to move all to https://www.digitalocean.com/) hosting biocache, collectory, communication portal, inhouse applications and soon also BIE and regions/spatial portal. This hosting is financed by the Guanacaste Conservation Area (https://www.acguanacaste.ac.cr/).

Australia - Amazon EC2, CSIRO data centre (VMWare), Nectar & NCI eResearch Infrastructure (OpenStack). Nectar is only used for non critical applications. For high availability public facing apps /tools we use Amazon (with load-balancing) but this is more expensive than the other 2 options we use. SOLR and Cassandra use is by far the most expensive. Expense is a concern and it is something we are actively trying to reduce.

	Running servers	**Total CPUs**	**Total memory (GB)**	**Total storage (GB)**
Amazon	68	261.52	1318	40334
Nectar	35	78	277	5,315
NCI	20	39	114	1740
CSIRO	17	68	937	11086

Philippines - Currently, the instance has been deployed in a VM instance of GCP. It strictly follows the minimum requirement as indicated in the ala-install.

- Ubuntu 16
- 4 vCPUs
- 15 GB memory

We have roughly estimated that the GCP could charge in as much as ~USD 2,088 per year if such VM would continuously be running. Relatively, this costs a lot as compared to an ordinary application. In addition, reconfiguring the system’s database separately from an instance to utilize the Cloud SQL of GCP is necessary for data back up. And this would add some additional charges. On its apparent state, adding CAS and customizing its look are being highly prioritized.

Portugal - See here. The cloud environment provided by INCD is managed with Openstack. The resources available in the project created in OpenStack are

- up to 30 instances
- up to 80 VCPUs
- up to 130 GB RAM
- up to 10 volumes

- up to 2 TB of storage

Now uses 6 instances, 42 VCPUs, 78 GB RAM, 4 volumes, 1.2 TB storage
New servers to be acquired by RI PORBIOTA will be integrated in INCD, under the agreement that part of hardware resources can be used by other INCD communities.

Guatemala: CONAP (Consejo Nacional de Áreas Protegidas), 1 VM hosted in google cloud with 16GB RAM, 100GB HB. (Waiting for the approval for resource to move to 3 VM), we have problems to costs the VM. (http://snidbgtbio.conap.gob.gt/)

Chile - we have a public portal called "Inventario Nacional de Especies", but we install the ALA portal Gbif Chile in demo mode, make and change configurations and desing (one VM server 16GB ram and 2TB disk) , also we want make deploy ALA in Docker, because all of system in the we work must be migrated to Docker technology. (http://especies.mma.gob.cl/CNMWeb/Web/WebCiudadana/Default.aspx)

Norway: Currently we have a dedicated (test) LivingAtlas VM hosted (and operated) by the university IT-department (Red Hat Enterprise 7.3, extra storage and memory compared to standard VMs). The university collaborates on hosting a "national" Infrastructure as a service (IaaS cloud service) that we could use. This IaaS uses OpenStack and supports Docker etc. The cost for the GBIF-node is moderate. Pain points include limited (but possible on VPN) ssh access from outside the university campus; limited access to provide access for users outside the University of Oslo - and even more problems for users outside the university network in Norway. We are exploring other possibilities, including setting up services in national cloud services directly on the same infrastructure as the university is using - possibly skipping some of the pains of security issues for ssh and user authorization - but probably creating new pain points. The cloud infrastructure will be managed - and GBIF-Norway would only operate the Living Atlas components themselves. All this is still under development!

Austria: Biodiversity Atlas Austria (available at www.biodiversityatlas.at, hopefully in autumn 2019); projects runs until end of 2020 - need to secure funds for afterwards
Server hosted by ZAMG (Central Institution of Meteorology and Geodynamics):
1VM, ubuntu 16.04 LTS, 8 CPUs, 64GB RAM, 5TB of storage;
As of now: server was set up and demo installed, next steps: configurations, etc…;

Russia: one aws ec2 instance + VPS on institutional server (Xen). Both - ubuntu 16.04

-----------------------------------End Infrastructure description-------------------------------------