Taxo&Map: procediment d'ingesta de dades i implementació

21/11/2023

Entorns

Desenvolupament

https://labs.geomatico.es/taxomap4

Producció

https://taxomap.bioexplora.cat

Proveïdors de dades

Els proveïdors lliuraran les dades en Simple Darwin Core[1] expressat en format de text tabulat.[2]

Es pot tractar d'un fitxer CSV, Excel (.xls), OpenDocument Spreadsheet (.ods), o bé d'un Darwin Core-Archive (DwC-A).

La primera fila del fitxer contindrà els noms de les columnes. Les columnes s'anomenaran amb el nomenclatura definida pels Darwin Core Terms.[3]

Taxo&Map accepta qualsevol camp definit per l'estàndard Darwin Core. No s'accepten camps que no estiguin definits en aquest estàndard.

Camps obligatoris

Els camps mínims requerits per l'aplicació són:

1http://rs.tdwg.org/dwc/terms/simple/

2http://rs.tdwg.org/dwc/terms/guides/text/index.htm

3http://rs.tdwg.org/dwc/terms/. Les correspondències entre els Darwin Core Terms i les diferents versions de Darwin Core es troben a http://rs.tdwg.org/dwc/terms/history/versions/.

Dades d'identificació:

- institutionCode
- collectionCode
- catalogNumber

El camp catalogNumber ha de ser únic per cada combinació de institució i col·lecció.

Dades taxonòmiques:

- scientificName
- kingdom
- phylum
- class
- order
- family
- genus
- specificEpithet
- infraspecificEpithet
- GBIF backbone identifier

S'estableix el GBIF Backbone (https://www.gbif.org/dataset/d7dddbf4-2cf0-4f39-9b2a-bb099caae36c) com a referència per a unificar les espècies dels diferents proveïdors de dades.

És tasca del proveïdor proporcionar els camps taxonòmics conforme a l'estàndard Darwin-Core i conforme a la taxonomia normalitzada del backbone de GBIF. Per tal d'automatizar la taxonomia, es recomana usar aquesta eina de matching: https://www.gbif.org/tools/species-lookup

Aquesta eina retorna un llistat amb els noms de les espècies, el nom científic vàlid, el seu identificador, la classificació taxonòmica, autoria, etc. i indica si la coincidència ha estat exacta o no.

Dades geogràfiques:

- decimalLatitude
- decimalLongitude

Si no es disposa de les dades geogràfiques serà necessari disposar de la informació de lloc completa que permeti fer una georeferenciació automàtica.

- country o countryCode
- stateProvince + county + municipality + locality

Cal tenir en compte que el camp locality només es podrà utilitzar per a la georeferenciació si conté topònims simples, però no si conté descripcions en text lliure del lloc de recol·lecció.

Dades temporals:

- year
- month
- day

Camps recomanables

Es recomana també proveïr com a mínim les següents dades:

- scientificNameAuthorship
- recordedBy
- eventDate
- identifiedBy
- dateIdentified
- minimumElevationInMeters
- maximumElevationInMeters
- minimumDepthInMeters
- maximumDepthInMeters
- coordinateUncertaintyInMeters

Preparació dels fitxers d'occurències

Vegeu també: Annex I: Modificacions al fitxer d'ocurrències del MZB i l'Herbari

Verificació dels noms de les columnes

- Verificar que els noms de les columnes utilitzen la nomenclatura definida per Darwin Core Terms.
- En cas contrari caldrà adaptar-los a Darwin Core[4] segons la taula de correspondències entre les diferents versions.[5]
- Els camps que no tinguin una correspondència directa amb Darwin Core s'eliminaran.
- Verificar que els valors de les columnes es corresponen amb el que indica l'estàndard.

Verificació dels requeriments mínims

- Verificar que el fitxer conté els camps mínims obligatoris.
- Si s'escau, els camps institutionCode i collectionCode es poden afegir massivament previ acord amb el proveïdor.
- Verificar que el camp catalogNumber no contingui duplicats dintre d'una mateixa col·lecció.
- Verificar la correspondència entre la classificació taxonòmica i el nom científic, en particular pel que fa a la presència dels camps genus, specificEpithet i infraspecificEpithet. En cas que aquests camps no hi siguin:
    - es pot mirar d'omplir-los automàticament separant els diferents components del nom científic
- Revisar si les ocurrències disposen de coordenades geogràfiques o de la informació mínima per obtenir-la.

Creació dels identificadors interns

Es crearan els següents identificadors per a ús exclusiu intern de l'aplicació:

- occurrenceID: les ocurrències s'identificaran amb la Darwin Core Triplet, que té la següent sintaxi: [institutionCode]:[collectionCode]:[catalogNumber]
    - P.e.MCNB:MCNB-Malac:MZB_2008-1246

4 http://rs.tdwg.org/dwc/terms/
5 http://rs.tdwg.org/dwc/terms/history/versions/

Annex I: Modificacions al fitxer d'ocurrències del MZB i l'Herbari

La següent taula resumeix les modificacions fetes al fitxer tot_utf8_original.csv. El resultat està al fitxer tot_utf8_modificat.ods.

En particular, s'han modificat els noms de les columnes per ajustar-los als termes de DarwinCore (segons http://rs.tdwg.org/dwc/terms/ i http://rs.tdwg.org/dwc/terms/history/versions/index.htm).

Tots els noms corresponents a termes de DwC s'han posat amb la primera lletra minúscula. Els canvis subsancials de nom s'han marcat en blau. Els camps nous en verd. En vermell, els camps a eliminar.

Nom anterior	**Nom actual**	**Observacions**
DateLastModified	modified	
InstitutionCode	institutionCode	Modificat MZB per MCNB per coherència amb el GBIF. De totes maneres, això s'haurà de modificar a l'origen de les dades.
CollectionCode	collectionCode	Modificats els valors. Al MZB varia segons la col·lecció. De totes maneres, això s'haurà de modificar a l'origen de les dades. NOTA: ho he canviat per a que tinguem en compte que no ens podem fiar massa d'aquest camp!
CatalogNumber	catalogNumber	S'han afegit els valors del MZB que estaven a l camp SpecimenId
ScientificName	scientificName	
BasisOfRecord	basisOfRecord	He completat registres de MZB amb "PreservedSpecimen"
Domain	domain	Camp propi
DomainId	domainId	Camp propi
Kingdom	kingdom	
KingdomId	kingdomId	Camp propi
Phylum	phylum	
PhylumId	phylumId	Camp propi
Class	class	
ClassId	classId	Camp propi
Order	order	
OrderId	orderId	Camp propi

Family	family	
FamiliyId	familyId	Camp propi
Genus	genus	
GenusId	genusId	Camp propi
Species	specificEpithet	
SpeciesId	speciesId	Camp propi
Subspecies	infraspecificEpithet	
SubspeciesId	subspeciesId	Camp propi
CanonicalName	canonicalName	Camp propi: defineix l'espècie i permet fer cerques a fots externes (Wikipedia)
ScientificNameAuthor	scientificNameAuthorship	
IdentifiedBy	identifiedBy	
	dateIdentified	Camp nou, unió dels 3 camps següents. =AH2 & IF(ISBLANK(AI2), "", "-" & IF(LEN(AI2)=1, "0","") & AI2 & IF(ISBLANK(AJ2), "", "-" & IF(LEN(AJ2)=1, "0","") & AJ2)) Format YYYY-MM-DD. http://en.wikipedia.org/wiki/ISO_8601#Dates Per a les cerques, es pot tractar simplement com un camp de tipus text (i no de tipus data)
YearIdentified		Eliminar
MonthIdentified		Eliminar
DayIdentified		Eliminar
TypeStatus	typeStatus	
CollectorNumber	recordNumber	
FieldNumber	fieldNumber	
Collector	recordedBy	
	eventDate	Camp nou format a partir dels 3 següents: =AH2 & IF(ISBLANK(AI2), "", "-" & IF(LEN(AI2)=1, "0","") & AI2 & IF(ISBLANK(AJ2), "", "-" & IF(LEN(AJ2)=1, "0","") & AJ2)) Format YYYY-MM-DD segons http://en.wikipedia.org/wiki/ISO_8601#Dates Per a les cerques, es pot tractar

		simplement com un camp de tipus text (i no de tipus data)
YearCollected	year	És redundant respecte el camp eventDate. En cas d'interval conservar el primer any.
MonthCollected	month	És redundant respecte el camp eventDate. En cas d'interval conservar el primer mes.
DayCollected	day	És redundant respecte el camp eventDate. En cas d'interval conservar el primer dia.
Country	countryCode	Tot i que hi ha algun error, en general es tracta del codi, no del país
StateProvince	stateProvince	
Locality	locality	
Longitude	decimalLongitude	
Latitude	decimalLatitude	
CoordinatePrecision	coordinateUncertaintyInMeters	
MinimumElevation	minimumElevationInMeters	
MaximumElevation	maximumElevationInMeters	
MinimumDepth	minimumDepthInMeters	
MaximumDepth	maximumDepthInMeters	
Sex	sex	
PreparationType	preparations	
IndividualCount	individualCount	
Notes	occurrenceRemarks	
SpecimenId		Eliminar aquest camp. Els valors que tenia el MZB s'han posat a catalogNumber

Annex II: Instruccions SQL per unificació de diferents proveïdors

Unificar CSV i importar dades dins de PostGIS mitjançant DBeaver.

```
/* decimal coordinates and uncertainty */

update dwc_merge set decimallongitude = lon, decimallatitude=lat, coordinateuncertaintyinmeters=CAST(opencage_radius_km AS numeric) * 1000

insert into mcnb_merge (select * from dwc_merge)


/* event date */

UPDATE  mcnb_merge  SET year = 9999 WHERE left(eventdate, 4) = '' OR substring(eventdate, 6, 2) = '' OR  substring(eventdate, 9, 2) = ''


UPDATE  mcnb_merge  SET year = 9999 WHERE left(eventdate, 4) LIKE '%-%' OR substring(eventdate, 6, 2) LIKE '%-%' OR  substring(eventdate, 9, 2) LIKE '%-%'


UPDATE  mcnb_merge  SET year = 9999 WHERE left(eventdate, 4) LIKE '%/%' OR substring(eventdate, 6, 2) LIKE '%/%' OR  substring(eventdate, 9, 2) LIKE '%/%'


UPDATE mcnb_merge SET year = CAST(left(eventdate, 4) AS integer), month = CAST(substring(eventdate, 6, 2) AS integer), day = CAST(substring(eventdate, 9, 2) AS integer) WHERE year IS NULL


UPDATE  mcnb_merge  SET year = NULL WHERE year = 9999


/* geometries */

UPDATE mcnb_prod SET geom = ST_SetSRID(ST_MakePoint(decimallongitude, decimallatitude),4326);
```

--i recarregar feature type a geoserver!

/* eliminar ocurrències sense camps taxonòmics obligatoris */

UPDATE mcnb_prod SET domain = 'Eukaryota'

UPDATE mcnb_prod SET kingdom='Animalia' WHERE kingdom = 'Animalia '

DELETE FROM mcnb_prod WHERE phylum IS NULL

/* trim: eliminar espais en blanc */

update mcnb_prod set phylum = trim(phylum)

update mcnb_prod set class = trim(class)

update mcnb_prod set class = trim(class)

update mcnb_prod set _order = trim(_order)

update mcnb_prod set family = trim(family)

update mcnb_prod set genus = trim(genus)

update mcnb_prod set species = trim(species)

update mcnb_prod set subspecies = trim(subspecies)

/* uniformitzar institutioncode i basisofrecord */

update mcnb_prod set institutioncode= 'Museu Valencià d''Història Natural' where institutioncode='MVHN'

update mcnb_prod set institutioncode= 'Universitat de Barcelona' where institutioncode='FEHM-UB'

update mcnb_prod set institutioncode= 'Museu Ciències Naturals Barcelona' where institutioncode='MCNB'

update mcnb_prod set basisofrecord= 'Fossil/Fòssil/Fósil' where basisofrecord='FossilSpecimen'

update mcnb_prod set basisofrecord= 'Non-fossil/No fòssil/No fósil' where basisofrecord='HumanObservation' OR basisofrecord='PreservedSpecimen'

/* errors UB */

update mcnb_prod set lon=decimallatitude where institutioncode='Universitat de Barcelona'

update mcnb_prod set decimallatitude=decimallongitude WHERE institutioncode='Universitat de Barcelona'

/* generar ids taxonomics */

UPDATE mcnb_prod SET species = genus || ' ' || specificepithet WHERE genus IS NOT NULL

UPDATE mcnb_prod SET subspecies = genus || ' ' || specificepithet || ' ' || infraspecificepithet WHERE species IS NOT NULL

/* indexar */

CREATE INDEX mcnb_prod_idx ON mcnb_prod USING GIST (geom)

/* camps temporals */

Aquests tres camps han de ser numèrics. En cas que es disposi només del camp eventDate sota el format 'yyyy/mm/dd', cal crear els camps year, month i day com a string i editar-los amb comandes SQL:

UPDATE mcnb_prod_sql SET year = left(eventdate, 4), month = substring(eventdate, 6, 2), day = substring(eventdate, 9, 2) WHERE institutioncode='Museu Ciències Naturals Barcelona'

Molt sovint la instrucció anterior pot fallar perquè no pot fer el CAST degut a errors. Es recomana:

UPDATE mcnb_merge SET year = 9999 WHERE left(eventdate, 4) = '' OR substring(eventdate, 6, 2) = '' OR substring(eventdate, 9, 2) = ''

UPDATE mcnb_merge SET year = 9999 WHERE left(eventdate, 4) LIKE '%-%' OR substring(eventdate, 6, 2) LIKE '%-%' OR substring(eventdate, 9, 2) LIKE '%-%'

UPDATE mcnb_merge SET year = 9999 WHERE left(eventdate, 4) LIKE '%/%' OR substring(eventdate, 6, 2) LIKE '%/%' OR substring(eventdate, 9, 2) LIKE '%/%'

UPDATE mcnb_merge SET year = CAST(left(eventdate, 4) AS integer), month = CAST(substring(eventdate, 6, 2) AS integer), day = CAST(substring(eventdate, 9, 2) AS integer) WHERE year IS NULL

```
UPDATE  mcnb_merge  SET year = NULL WHERE year = 9999

/*identificadors interns */

UPDATE mcnb_prod SET occurrenceid=institutioncode || ':' || collectioncode || ':' || catalognumber WHERE occurrenceid IS NULL
```