Fer accessible la informació científica: l'experiència del Museu de Ciències Naturals de Barcelona amb el domini Bioexplora

Francesc Uribe, conservador d'invertebrats no artròpodes (MCNB)

Jordi Agulló, gestió científica (MCNB)

Eulàlia Garcia Franquesa, cap de Col·leccions (MCNB)

CONTEXT

El concepte *big data* està plenament assentat. Com també està assimilat l'efecte multiplicatiu que suposa la capil·laritat del *linked data*. Al mateix temps una constel·lació de criteris de socialització de la informació, principalment *open data*, *open access*, etc., aparionen un vastíssim univers de persones usuàries. Parlem doncs de disponibilitat d'immenses bosses d'informació al servei d'enormes quantitats de potencials usuaris guarnits d'heterogeneïtat creixent en la mesura que creix l'accés a les fonts d'informació.

Des del sector de proveïdors d'informació de ciències naturals la nostra posició és manifestament a favor dels estàndards documentals, sigui en entorns manifestament naturalistes, sigui quan aquests productes es comuniquen amb altres sectors de la informació. Afortunadament, la comunitat de professionals que es reconeixen com a informàtics de la biodiversitat ha treballat amb molt esforç per produir esquemes documentals d'ampli ressò.

Una organització exemplar, el TDWG, acrònim de *Taxonomic Databases Working Group* ara més conegut com a *Biodiversity Information Standards*[1], ha conduït l'elaboració d'estàndards per a l'intercanvi de dades de biologia en general i de biodiversitat de forma més concreta. El més conegut i reconegut és *Darwin Core*, amb un afinat diàleg amb *Dublin Core*, però també destaquen *Natural Collections Descriptions* adreçat a descriure conjunts d'elements, *Audobon Core* per a recursos multimèdia i altres. Recentment s'hi ha sumat el Global *Genome Biodiversity Network* fruit d'una voluntat compartida entre el TDWG i la GGBN per

1 http://www.tdwg.org/

nodrir d'eines de comunicació de dades viables i etiquetades en entorns d'agregació de múltiples fonts, i per la creació del GGBN Data Standard[2] que incorpora vocabularis controlats de termes molecular als estàndards *Darwin Core* i *Access to Biological Collections Data* [3].

A la base del desplegament dels serveis web de dades patrimonials proveïdes pel Museu de Ciències Naturals de Barcelona (MCNB) hi ha el concurs potent i rigorós dels estàndards documentals. Aquesta base permet activar la nostra contribució de dades genuïnes del Museu en un entorn propi així com en amplis repositoris mundials de dades[4].

La disseminació de dades a partir d'eines web pròpies i la càrrega d'informació del MCNB a portals d'integració es beneficien de compartir l'aplicació d'estàndards i d'esquemes de dades. Internament, això es tradueix en que l'explotació d'una mateixa font de dades del museu pot preveure nodrir un servei propi de difusió al temps que recorre la passarel·la que connecti amb una federació de dades compartides. El principi de COPE, *Create Once and Publish Everywhere*[5], té un valor estratègic. O potser són dos valors. L'un és de l'economia de recursos per mobilitzar la informació i l'altre és l'evidència de que la informació sistematitzada afegeix qualitat documental a la quantitat de dades.

ELS PROJECTES DEL MCNB AMB MÉS DETALL, EL CAMÍ SEGUIT I ELS RESULTATS

A partir dels anys 80, al Museu de Zoologia de Barcelona (MZB, ara integrat al MCNB) dicta i segueix recomanacions per a la documentació de col·leccions de ciències naturals. Aquestes es transformen en les directrius que van guiar els anys 90 l'ús del programari Documentació Assistida de Col·leccions (DAC) per aplicació de la Llei de Museus i el decret 35/1992, fins a l'actualitat, amb la guia de documentació de les col·leccions zoològiques[6]. La documentació i la informatització de les col·leccions ha descansat des de l'inici sobre els

2 http://wiki.ggbn.org/ggbn/GGBN_Data_Standard_v1
3 http://data.ggbn.org/schemas/ggbn/Enviro/ABCDGGBN_Enviro.html
4 https://www.gbif.org/
5 P.e. https://collectionstrust.org.uk/resource/create-once-publish-everywhere-cope/
6 https://museuciencies.cat/area-cientifica/colleccions/documentacio-de-colleccions/

vocabularis controlats (tesaurus, diccionaris...). En el curt termini es disposarà també de guies finalitzades per a les col·leccions geològiques, paleontològiques i botàniques.

Assumit l'ús dels estàndards, les dades comencen a ser publicades el 2007 al portal internacional *Global Biodiversity Information Facility*, GBIF[7]. Es tracta d'una infraestructura intergovernamental que comprèn en l'actualitat 53 països i 43 organitzacions internacionals. GBIF s'estructura en nodes nacionals amb l'objectiu de donar accés via internet a les dades de biodiversitat de tot el món seguint l'estàndard de documentació *Darwin Core*. El MCNB actualitza anualment les seves dades, en aquests moments disposa de 120.548[8] registres publicats, de les col·leccions botàniques, zoològiques i paleontològiques.

Les dades de la col·lecció de vertebrats (col·lecció de Cordats) es poden consultar també a VerNet[9]. Es tracta d'una plataforma d'agregació de col·leccions originalment de vertebrats que aplega així mateix professionals en la gestió de dades de biodiversitat i la informàtica de la biodiversitat. VertNet lidera diversos projectes mundials de bases de dades de biodiversitat[10]

A més de la contribució neta en repositoris mundials, el MCNB ha desenvolupat eines web pròpies per a la disseminació dels continguts creats al museu. Aquestes aplicacions en van concentrar l'any 2009 en un domini comú creat amb l'objectiu principal de poder incorporar tecnologies de base en ràpida evolució: el portal Bioexplora (http://www.bioexplora.cat/). Aquest espai web, més experimental i oportunista es combina i sincronitza amb el web de comunicació oficial del museu.

7 https://www.gbif.org/
8 https://www.gbif.org/publisher/e8eada63-4a33-44aa-b2fd-4f71efb222a0
9 http://portal.vertnet.org/p/museu-de-cincies-naturals-de-barcelona
10 http://www.vertnet.org/about/partners.html#t-tab4

A Bioexplora s'hi destaquen projectes com:

Georeferenciació és l'expressió pública del treball de documentació interna que permet georeferenciar les localitats de recol·lecció de les col·leccions. Abans dels dispositius de geolocalització la ubicació en un mapa era una informació freqüentment textual: assignar-hi coordenades que permetin transportar els orígens de recol·lecció a mapes digitals és una tasca tècnica de gran importància. Transparentar els nostres resultats és un servei afegit que Georef proporciona a altres entitats.

Taxo&map representa l'estratègia de cerca complementària a l'habitual dels formularis. A **Taxo&map** la cerca s'inicia amb la informació visible i l'acció comença filtrant per categories taxonòmiques i límits geogràfics. El següent pas serà incorporar una línia de temps. **Taxo&map** és avui en dia una aplicació compartida amb altres museus. Un objectiu destacat és que ofereix molts formats per exportar les dades.

Wikicollecta té un paper estratègic per generar projectes diversos a partir d'una tecnologia d'edició gairebé autònoma als tècnics del museu. Ara són visibles el projecte sobre **Espècimens tipus** (exemplars lligats a la descripció científica de noves espècies) i **Protagonistes** (inventari de persones i institucions vinculades a la història del museu). A la rebotiga s'estan preparant altres projectes i algun ja ha volat com és el cas del web **Georeferenciació** que ha adquirit vida pròpia.

Catàleg 3D és una col·lecció d'imatges en tres dimensions generades a partir d'exemplars del fons de col·leccions del Museu. Actualment acull un projecte, l'Atles osteològic 3D, que permet la visualització en 3D d'elements ossis d'esquelets de la col·lecció de vertebrats. Les imatges s'acompanyen d'informació detallada. Es mostra doncs, una part de la col·lecció

científica, habitualment no exposada. Destacar que la presa de mesures sobre la imatge s'ha mostrat perfectament vàlida en estudis científics[11].

Guia de Fons i col·leccions. La Guia (http://www.bioexplora.cat/ncd/home/lang-nl) és una eina per a posar a l'abast de ciutadans i investigadors les col·leccions del MCNB i de l'IBB a partir de descripcions de col·leccions. Les fitxes segueixen l'esquema de metadades *Natural Collections Description.* Aquesta descripció per a conjunts de registres significatius es coneix com a *Collections-Level Description*. L'enfoc documental té el seu origen en la descripció arxivística, quan als anys 90, la *Society of American Archivist* y la *Library of Congress*, dissenyaren un esquema de metadades per a la codificació dels instruments de descripció d'arxius (*finding aids*) en l'entorn digital: el *Encoded Archival Description* (EAD)[12]. S'han elaborat unes 240 fitxes de descripció de col·leccions i una trentena de fitxes de departament o àrea.

Col·leccions obertes (http://www.bioexplora.cat/ca/colleccions-obertes), actualment en desenvolupament, es tracta d'un magatzem de dades al núvol amb el model de dades *Resource Description Framework* (RDF) i diferents esquemes de metadades, que es nodreix de les dades de col·leccions actualment en diferents bases de dades (Access, Museumplus i Filemaker). La funcionalitat principal del magatzem de dades és permetre fer consultes i recuperar els resultats amb un model de dades amigable per als desenvolupadors de webs. L'objectiu és poder mostrar un catàleg de l'inventari del Museu de totes les col·leccions que formen el MCNB. A més, es vol aconseguir una aproximació a les col·leccions del Museu d'una forma visual, a partir de la seva representació gràfica, fent-la més amena a la seva descoberta.

---

[11] Quesada, J., Aurell-Garrido, J., Gago, S., Boet, O & Garcia-Franquesa, E. 2016. Measurements errors in 3D models used in osteometric data research with freeware: a test using skulls of the Algerian hedgehog (*Atelerix algirus*). *Vertebrate Zoology* , 66 (3): 411-418.

12 http://www.loc.gov/ead/

Les **Publicacions científiques** del museu també formen part de **Bioexplora**. Pel conjunt d'aplicacions integrants de **Bioexplora** el portal ofereix una base tecnològica dinàmica, serveis comuns a les xarxes socials i especialment cal destacar la voluntat de respondre a un repte molt significat: crear una comunitat de suport digital als continguts publicats pel museu. Aquest suport comença per obrir la possibilitat de que els visitants puguin traslladar comentaris en relació a la gran majoria de pàgines web. El següent pas serà establir projectes explícitament adreçats al voluntariat digital. La col·laboració entre usuaris i responsables tècnics del museu persegueix la detecció d'errors o de complements a les dades aportades.

FUTUR

Interoperativitat

El concepte d' *open linked data* descriu amb propietat la vinculació de les dades del museu amb els repositoris internacionals. Una mirada interior obre noves perspectives, les de combinar, sumar, enllaçar les diverses fonts d'informació inèdita que representen el patrimoni intel·lectual del museu. En la pròpia lògica d'evolució de **Bioexplora,** un punt de concentració de serveis de dades, emergeix un objectiu a curt termini: establir una capa d'interoperativitat que construeixi ponts semàntics i de dades entre els orígens d'informació responsabilitat del museu.

Aquesta visió ja té exemples d'aplicació entre alguns dels projectes actius: del **catàleg 3D** i **col·leccions obertes**, **Wikicollecta** i **Taxo&map**. El propòsit actual és crear un marc de connexions i reutilitzacions entre fonts idoni per establir complementarietats documentals i per gestionar les possibles disparitats entre vocabularis compartits, un exercici delicat però sovint higiènic.

No acaba aquí el sentit d'una capa on circuli la informació entre fonts: també ofereix una base neta per a projectes d'alt nivell d'integració de dades. Un exemple és explicar el relat

de les expedicions i excursions científiques del museu, al qual hi concorren col·leccions, arxiu, publicacions, mapes i persones.

Qualitat de la informació

L'acumulació de continguts suposa un major risc de perdre perspectiva de l'origen de les dades, la qual cosa significa un increment en part descontrolat d'acceptar informació de baixa qualitat. És evident que els continguts estandarditzats contenen les condicions per a una depuració orientada. L'ampliació de la informació generada pel Museu s'hauria de portar a terme tenint present l'ús d'eines d'avaluació de la qualitat de les dades quan aquestes són publicades. Existeixen eines per la neteja i validació de les dades, com per exemple: *Darwin Test*[13], *Kurator*[14], *DarwinCore Archive Validator*[15], *OpenRefine*[16] etc.

Però no n'hi ha prou i per aquí s'espera que es produeixin els moviments de futur per induir avaluacions de la qualitat de les dades. En aquest sentit els primers passos són permetre la participació de la comunitat d'usuaris en defectes puntuals d'informació.

Des d'un punt de vista més estructural, la participació en portals de dades sovint significa adquirir metadades de les col·leccions o directament indicadors o mètriques que apunten a mancances d'informació. El terreny està obert a sistematitzar l'aplicació d'eines de depuració de dades a l'origen dels mateixos proveïdors d'informació. Un esforç que podria ser reconegut per alguna distinció de qualitat de la font.

Ontologies

Un altre factor de canvi serà la incorporació d'ontologies com a referència dels camps que produeixen molta terminologia amb múltiples vinculacions entre termes. Davant d'aquesta

13 http://www.gbif.es/software/darwin-test/

14 http://wiki.datakurator.org/wiki/

15 https://tools.gbif.org/dwca-validator/

16 http://openrefine.org/

situació els vocabularis controlats clàssics no aconsegueixen resoldre la documentació original, per bé que la persona que consulti pugui tenir la il·lusió de que troba una informació molt adient. Les ontologies en ciències naturals són eines implicades en els processos de recerca, per exemple en la descripció anatòmica[17], i suposen ja un conjunt ampli de referències assimilables[18].

Per raó de l'origen especialitzat d'aquestes ontologies, els gestors d'informació dels museus de ciències naturals han de trobar el compromís entre granularitat de la descripció que ofereixen les ontologies i la textura de la informació disponible, ocasionalment més difusa i heterodoxa del que es troben els investigadors al laboratori.

EPÍLEG

El MCNB disposa d'informació científica generada des del fons patrimonial així com des de l'activitat científica dels tècnics de col·leccions. Aquesta informació es fa pública a mesura que es va generant, acomplint una de les funcions bàsiques d'un museu públic, divulgar i compartir amb la ciutadania. Els projectes actuals permeten mostrar diferents disciplines de les ciències naturals. Els projectes nous integraran encara més el conjunt de la informació generada al Museu (col·leccions, publicacions, documents d'arxiu, recerca, activitats, etc.).

L'empenta i esforç dels tècnics de col·leccions, juntament amb les eines tecnològiques existents, fan que aquest Museu es trobi plenament integrat en el segle XXI. L'experiència viscuda juntament amb els aprenentatges acumulats garanteixen que en el futur la visualització del patrimoni científic pugui tenir encara més valor per a projectes de geo i biodiversitat d'arreu, projectes naturalistes, culturals i de ciència ciutadana.

17 P.e. Uberon https://uberon.github.io/
18 http://www.ontobee.org/