

Cap on va Invenio 3?

Conclusions de l'*Invenio v3 Bootcamp*

CERN, 19-21 March 2019

Ferran Jorba, UAB, 3 abril 2019

Sumari

- Com va anar la trobada al CERN
- Continuen els canvis de personal
- Objectius d'Invenio 3
- Requisits informàtics
- Temes que encara estan pendents en Invenio 3
- Conclusions
- Propostes de futur

L'equip d'ara i la vella escola



Lars Holm Nielsen, Nicola Tarocco, Alexandros Ioannidis, Diego Rodríguez Rodríguez, Zacharias Zacharodimos i Karolina Przerwa, l'equip Invenio 3.



Alexander Wagner (DESY, Deutsches Elektronen-Synchrotron) i Tibor Šimko (CERN, autor d'Invenio 0.x i 1.x), els nostres socis Marc21.

Feina i vida social



35 persones, 35 portàtils, 35 mòbils i un munt de cables i ratolins...



Makoto Sumiyoshi (o és en Masaharu Hayashi?),
Lars Holm Nielsen i José Benito González López
(el cap de'n Lars)

Continuen els canvis de personal al CERN

- Ja no queda ningú de l'equip original. Ningú.
- José Benito González López continua essent el coordinador.
- Lars Holm Nielsen és el responsable de Zenodo i Invenio.
- Les presentacions ens les van fer un danès (Lars Holm Nielsen), dos espanyols (José Benito González López i Diego Rodríguez Rodríguez) dos grecs (Alexandros Ioannidis i Zacharias Zacharodimos), un italià (Nicola Tarocco) i una polonesa (Karolina Przerwa).
- Tots molt jovenets i entusiasmats amb les noves tendències informàtiques.
- A les primeres trobades d'Invenio hi havia alguna bibliotecària del CERN. Ara ja no en veig cap. Això deu tenir alguna conseqüència...

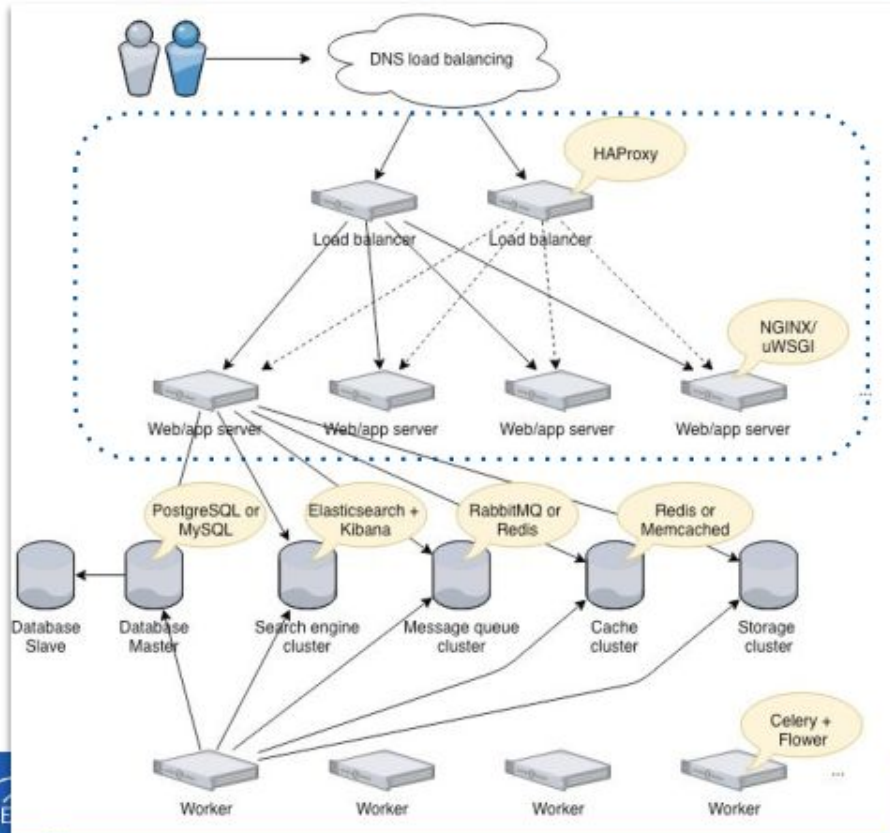
Què és (ara) Invenio 3?

- Una plataforma (*framework*) per desenvolupar repositoris.
- Deixa de ser un software *instal·lable*. Per fer-ne alguna cosa cal desenvolupar software en Python i seguir les convencions Flask.
- El que (segurament) serà instal·lable seran el que ells anomenen els *flavours*: Zenodo (tot i que no té sentit fora del Cern), RDM (Research Data Management, un futurible), etc.
- La llicència ha passat de GPL 2.0 (viral, que obliga que les obres derivades continuïn sent lliures) a MIT (que permet obres derivades tancades i propietàries).

Requisits informàtics d'Invenio 3

- Imprescindibles:
 - Python, amb Flask, pyenv i uWSGI, i mil altres paquets Python.
 - Servidor web: Nginx o Apache.
 - Servidor SQL: PostgreSQL o MariaDB (i SQLite per proves).
 - Indexador: Elasticsearch v5 o v6 (canvien de dalt a baix cada cop).
 - Gestor de tasques: Celery (que requereix RabbitMQ o Redis per emmagatzemar-les).
- Opcionals:
 - Gestor de sessions i caché: Redis o Memcached.
 - Balancejador i/o gestor d'alta disponibilitat: HAproxy.
 - Empaquetador d'aplicacions: Docker.

Infrastructure overview



- *Load balancers:* HAProxy, Nginx or others.
- *Web servers:* Nginx, Apache or others.
- *Application servers:* uWSGI, Gunicorn or mod_wsgi.
- *Distributed task queue:* Celery
- *Database:* PostgreSQL, MySQL or SQLite.
- *Search engine:* Elasticsearch (v5 and v6).
- *Message queue:* RabbitMQ, Redis or Amazon SQS.
- *Cache system:* Redis or Memcache.
- *Storage system:* Local, S3, XRootD, WebDAV and more.

Invenio 3, pendent (I): registres bibliogràfics

- Marc21 (amb el mantra: 'No us el recomanem')
 - No hi tenen cap interès: ho deixen a les nostres mans.
- Formularis
 - Ells mateixos reconeixen que està molt verd.
- Editor de registres bibliogràfics
 - Oficialment depèn de l'esquema en JSON del format bibliogràfic.
- Càrrega de registres
 - D'això ara en diuen *loaders*.
- Formats de sortida
 - Ara en diuen *serializers*.
- Indexació
 - S'ha de configurar l'ElasticSearch, que canvia cada versió.

Invenio 3, pendent (II)

- Col·leccions
 - No hi tenen cap interès: ho deixen a les nostres mans.
- Gestió de fitxers
 - No està implementat.
 - Ara per ara (ElasticSearch) no indexa el text complet dels fitxers.
- Usuaris, rols i permisos
 - No està clar que puguem tenir dos tipus d'autenticacions, LDAP pels interns i email pels externs.
 - El que vam veure semblava molt embrionari.
- OAI, servidor i client
 - No sembla que estigui integrat.

Aleshores, a què s'han dedicat, mentrestant?

Bona pregunta.

Si hem de fer cas en què s'entretienien més parlant a les presentacions...

AngularJS to React

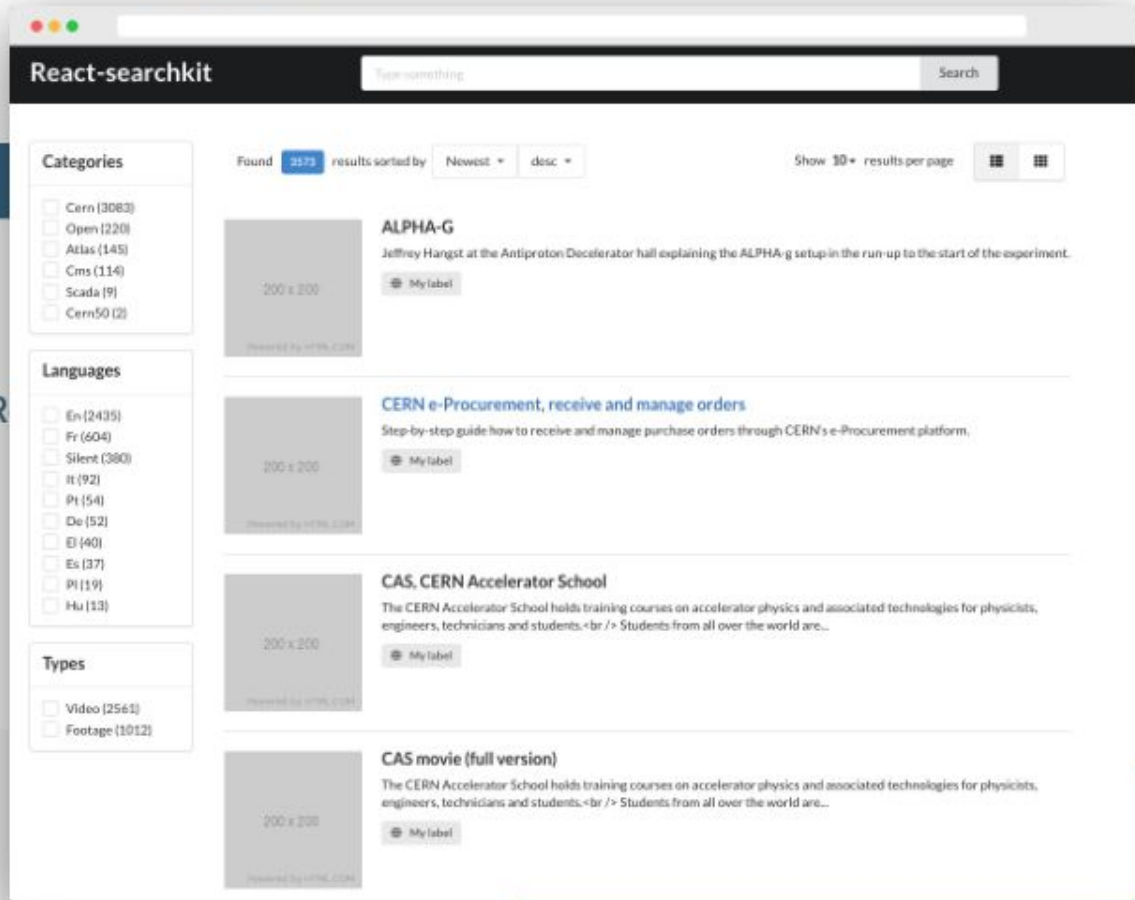
React-SearchKit

A simple yet powerful UI search kit built with R

GET STARTED

DISCOVER COMPONENTS

★ Star 12



Walk-through Invenio headers

- Secure defaults
- Mainly set by Flask-Talisman
- Still configurable in a per route basis
- Some more considerations...

```
HTTP/1.0 200 OK
Content-Length: 293
Content-Security-Policy: default-src 'self'; object-src 'none'
Content-Type: application/json
Date: Wed, 13 Mar 2019 21:38:36 GMT
Link: <https://127.0.0.1:5000/api/records/?pag...
Referrer-Policy: strict-origin-when-cross-origin
Retry-After: 3582
Server: Werkzeug/0.14.1 Python/3.6.7
Strict-Transport-Security: max-age=31556926; includeSubDomains
X-Content-Security-Policy: default-src 'self'; object-src 'none'
X-Content-Type-Options: nosniff
X-Frame-Options: sameorigin
X-RateLimit-Limit: 5000
X-RateLimit-Remaining: 4997
X-RateLimit-Reset: 1552516699
X-XSS-Protection: 1; mode=block
```

Presentacions rellevants (I)

- José Benito González López (CERN), *Welcome*
 - <https://indico.cern.ch/event/773969/contributions/3351793/>
 - Presentació general, història d'Invenio i projectes que l'utilitzen al Cern.
- Karolina Przerwa, *Anatomy of a repository*
 - <https://indico.cern.ch/event/773969/contributions/3351830/>
 - Un tast del que vol dir instal·lar Invenio 3
- Zacharias Zacharodimos, *Data models: Add a new field*
 - <https://indico.cern.ch/event/773969/contributions/3351834/>
 - Terminologia, conceptes i quina por...
- Nicola Tarocco, *Build a simple deposit form*
 - <https://indico.cern.ch/event/773969/contributions/3351843/>
 - Més por...

Presentacions rellevants (II)

- Karolina Przerwa, *Managing access to records*
 - <https://indico.cern.ch/event/773969/contributions/3351844/>
 - 100% Python.
- Diego Rodríguez Rodríguez, *Securing your Invenio instance*
 - <https://indico.cern.ch/event/773969/contributions/3351845/>
 - Com defensar-se d'atacs via capçaleres http, tokens i cookies.
- Nicola Tarocco, *Deployment and monitoring*
 - <https://indico.cern.ch/event/773969/contributions/3351846/>
 - La més completa des del punt de vista informàtic d'Invenio 3.
- Lars Holm Nielsen, *Roadmap and Invenio development @ CERN*
 - <https://indico.cern.ch/event/773969/contributions/3351852/>
 - Les peces que falten, (només) des del punt de vista informàtic.

Invenio 3, conclusions

- Encara estem lluny del que dóna de sí Invenio 1.x.
- La idea d'Invenio 3 era agafar peces ja desenvolupades per avançar més ràpid, però no sembla que ho estiguin aconseguint.
- Tind, en estat d'hivernació?
- Les traduccions, encallades.
- Molt JavaScript. Vaig preguntar al Lars com afecta, per exemple, les persones amb problemes visuals que utilitzen navegadors sense JavaScript, com el Lynx. O també els robots casolans. Es va quedar sorprès i van dir que no havien tingut temps de tractar aquest cas.

Invenio 3, dues opinions personals

- Informàtics en estat d'autocontemplació:
 - En cap moment van parlar d'autors, títols o publicacions periòdiques.
 - Ni tan sols acceptaven que la casella de cerca tingués un desplegable per camps, perquè diuen que depèn del model de dades.
 - Això sí: frameworks Python i Javascript, capçaleres http, alta disponibilitat, contenidors, càrrega distribuïda,...
 - Nosaltres: 'Us quedareu sols'. Cern: 'Doncs si ha de ser-ho, endavant'.
- L'evolució d'Invenio 1.x a Invenio 3, com la l'Omeka Classic a Omeka S:
 - Omeka Classic s'entén, Omeka S no.
 - S'obvien els conceptes obvis i consolidats i s'entra en mons abstractes de semàntica i conceptualització.

I doncs? Què fem a la UAB?

- Seguir la via Invenio 3?
 - Estem en el mateix vaixell que els alemanys (contacte: Alexander Wagner).
 - Ens van proposar formar part del consorci Join2 (<https://github.com/join2>) per avaluar-ho conjuntament.
 - Proposta de trobada a Hamburg el juny per l'OpenRepositories.
 - Es tractaria de crear un *flavour* Invenio 3 equivalent, en funcionalitat, a Invenio 1.x, instal·lable i configurable.
- Hi ha alternativa?
 - Tornem a mirar el que hi ha disponible?

Jo tinc una proposta

Us l'explico?

L'alternativa Muscat (I)

Un programari lliure, basat en Ruby on Rails, que gestiona:

- Marc21, incloent bibliogràfic, autoritats i holdings.
- Multilingüe (anglès, francès, italià i alemany; castellà i portuguès en marxa).
- Indexació a text complet (via Solr), inclou els documents adjunts.
- Control de versions dels registres integrat.
- Usuaris, rols, permisos, alertes, etc.
- Múltiples formularis per base de dades, amb camps amb autocompletesa.
- Amb geovisualització integrada.
- Especialitzada en música.
- <http://muscat-project.org/>

L'alternativa Muscat (II)

- Públic des del 2016.
- Desenvolupament: <https://github.com/rism-ch/muscat>, ara a la 5.1
- Amb col·leccions, i jeràrquiques? <http://muscat-project.org/model.html>
- Es poden crear carpetes per organitzar el treball.
- Possibilitat de deixar comentaris als registres bibliogràfics i d'autoritats, per revisar-los o treballar-hi conjuntament.
- Gestió d'objectes digitals (?),
- Els registres d'autoritat es poden lligar al VIAF per importar-ne els noms (potser també a Orcid?).
- Gestions específiques per música (ex: incipits)
- Programari lliure (licència no explícita).

Muscat: tutorials, presentacions i vídeos

- <http://www.rism.info/community/muscat.html>
 - Informació general, orientat sobretot als catalogadors de música.
- <http://www.rism.info/en/publications/iaml-congresses/2018.html>
 - Algunes presentacions interessants.
- <https://github.com/rism-ch/muscat/wiki>
 - Notes de development.
- https://www.youtube.com/watch?v=ncnQ-TD9dGM&list=PL9SyOIE9iSYLn timerJz3fGPkl8Xaf_ikDe4&index=3
 - Quatre vídeos (cal començar pel tercer) que expliquen les interioritats, especialment com funcionen les autoritats.
 - Pel que es veu al peu de pantalla, es tracta de la versió 3.5.2, i ara anem per la 5.1, de manera que algunes coses deuen haver millorat.

Exemples amb Muscat

- Exemple real: <http://rism-ch.org/>. 84.386 registres.
 - Registre amb fitxers adjunts: <http://www.rism-ch.org/catalog/402004752>
- L'altre: <https://www.canons.org.au/>, 2.216 registres
- Servidor de demo: <http://demo.muscat-project.org>.
 - 1.125,000 registres (*musical sources*), 111.700 registres d'autoritat de persones, 70.400 d'institucions, and 34.500 records de literatura secundària: total, 1.248.000 registres.
- Servidor de pràctiques: <https://muscat-training.rism.info>
 - Sembla que té els mateixos registres que l'anterior.
 - Però podem entrar a la zona interna:
 - Usuari: training01@rism.info fins a training99@rism.info
 - Contrasenya: password

Muscat també es basa en peces preexistents (I)

- Ruby on Rails, plataforma de desenvolupament web (<https://rubyonrails.org/>)
 - Força popular. Nosaltres en coneixem el Redmine.
- Solr, l'indexador (<http://lucene.apache.org/solr/>)
 - Després parlarem de la relació entre Solr i Elasticsearch.
- Ruby Marc (<https://github.com/ruby-marc/ruby-marc>)
 - En continu desenvolupament, sembla molt complet.
- Blacklight (<http://projectblacklight.org/>)
 - “An open source Solr user interface discovery platform”. Usat en diferents aplicacions bibliotecàries.
- Blacklight Marc (<https://github.com/projectblacklight/blacklight-marc>)
 - Millores de Blacklight específiques per a registres Marc.

Muscat també es basa en peces preexistents (II)

Gestió d'usuaris i permisos (<http://muscat-project.org/users-and-feedback.html>):

- Autenticació basada en Devise (<https://github.com/plataformatec/devise>)
 - Segons sembla, hauria de permetre múltiples tipus d'autenticació.
- Gestió de rols basada en Rolify (<http://rolifycommunity.github.io/rolify/>).
- Autoritzacions via CanCan (<https://github.com/ryanb/cancan>).

Geovisualització:

- Dariah GeoBrowser (<https://geobrowser.de.dariah.eu/>)

És a dir, se n'han ensortit amb el que Invenio 3 sembla que no.

Lucene, Solr, i Elasticsearch

Quina relació tenen entre sí? (el que he après, sense coneixements pràctics)

- Lucene (1999) és una Api en Java per crear indexadors. Tothom la lloa.
- Solr (2004) és un indexador basat en Lucene, i ara integrat al projecte.
- Elasticsearch (2010) és un altre indexador basat en Lucene, més enfocat al JSON i amb un biaix més comercial.
- Sembla haver-hi un consens en que són molt comparables (ex: <http://solr-vs-elasticsearch.com/>), en funcionalitat, rendiment i ambició.
- Tots dos coexisteixen i Internet està tan ple d'arguments a favor de l'un com de l'altre.

Invenio 3 o Muscat 5?

Avaluació provisional de la seva adequació i el nostre esforç d'implementació

- IFMuC
 - Invenio 3: 10% (massa gran, complexe i immadur).
 - Muscat: 90% (100% Marc21 i 100% música, caldria adaptar la interfície).
- Traces
 - Invenio 3: 20% (també massa gran, complexe i immadur).
 - Muscat: 80% (100% Marc21, caldria adaptar la interfície i formulari).
- DDD
 - Invenio 3: 30% (encara massa gran, complexe i immadur).
 - Muscat: 70% (100% Marc21, adaptar interfície, formularis i permisos).

Què li falta a Muscat?

Coses que hem de resoldre si el volem adoptar:

- Traducció al català (exemples de les altres traduccions: <https://github.com/rism-ch/muscat/tree/master/config/locales>).
- Servidor OAI (però: <https://github.com/code4lib/ruby-oai>).
- (Client OAI, tot i que no cal que estigui integrat; però: <https://github.com/fieldhand/fieldhand>).
- Col·leccions, o URLs compatibles amb les d'Invenio.
- Més comunitat, més instal·lacions? Segons com pot ser una avantatge, perquè et poden tractar millor.

És a dir, no són gaire diferents de les d'Invenio 3.

Proposta estratègica

1. Avaluem primer amb detall l'adequació de Muscat per IFMuC o Traces (voluntaris?).
2. Instal·lem-lo, provem-lo i aprenem-ne.
3. Dissenyem, traduïm i adaptem.
4. Si ho veiem bé, fem el pas a producció.
5. Repetir-ho per l'altre (Traces o IFMuC).
6. Revisem com ha evolucionat Invenio 3 mentrestant.
7. Avaluem l'adequació de Muscat per al DDD.
8. Decidim.
9. Fem.