Indico & Citadel Search: A collaboration case study

Pedro Lourenço, Carina Antunes (CERN)

OSSYM 2021 - 3rd International Open Search Symposium

2021/10/13

Fundamentals

Indico

- Open-source event management system, popular in the HEP (High-Energy Physics) community.
- Extensible through its plugin architecture (Paypal, video-conferencing, search engines, among others).
- Over 8M indexable records at CERN.

Lacked out-of-the-box search capabilities.

For many years, instances relied on locally-developed custom search strategies or services, often based on paid enterprise solutions, to fulfil this gap.
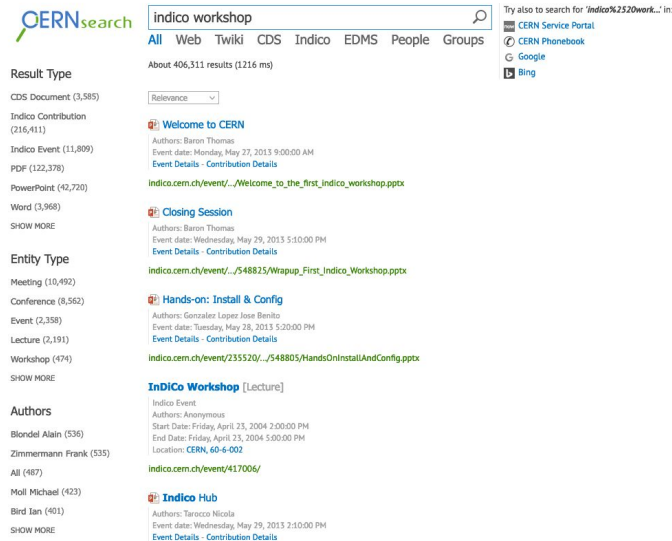
Fundamentals

Which engines were available:

- CERN Search - Sharepoint based
  https://search.cern.ch
- Invenio https://invenio-software.org

How were they integrated?

- Redirect to search engine's site (e.g. Invenio)
- Inside Indico: Retrieved by Indico (e.g. Invenio)
- Inside Indico: Embedded Iframe (e.g. CERN Search - Sharepoint based)

Fundamentals

Indico v1-2

- Proprietary search implementation based on Sharepoint Search.
- Relived the complexity of maintaining an external search engine.
- The interface is external - there is no native integration.
- The framework cannot be used outside of CERN.

SQL → Sharepoint → Citadel

The solution

Indico v3

- A new search component backed by *ElasticSearch*, sending its metadata in a JSON format compatible with existing database-linked schemas.

- Out-of-the-box native interface that contributes to a seamless integration of multiple search adapters.

- Can be easily replicated across external instances.

Citadel Search

Open Source enterprise search solution

State of the art technologies

Affordable & Transparent

Gives control to the users

- Enables users to create tailored data models for different information sources.

- Have fine grained access control over their information.

- Obtain relevant results upon search queries, by customizing scoring, indexing and document retrieval.

- Structured Data and Digital Content.

Citadel Search - Architecture

Stack

- Elasticsearch
- Tika
- Invenio
- Celery
- PostgreSQL
- Nginx
- RabbitMQ
- Ceph
- Redis
- Sentry
- OKD
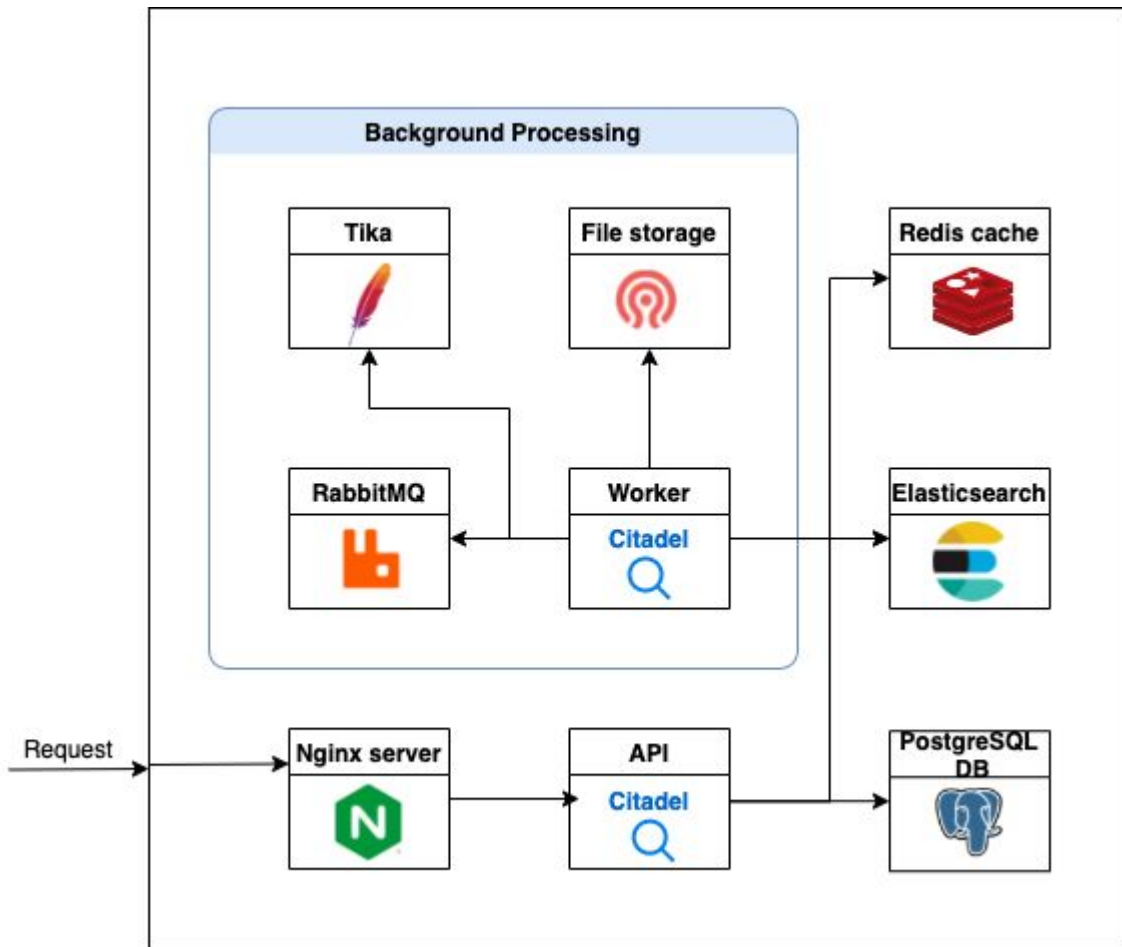
Citadel Search - Feature Overview

- **RESTful API**
  - CRUD Documents and Files, search via ES query strings syntax.

- **Abstracts and automates many of the burdens of dealing with Elasticsearch by HTTP and CLI actions**
  - create index, put mapping, reindex ...

- **Relational and non-relational storage**
  - Postgres for history and versioning of documents, allows recovery and reindexing.
  - Elasticsearch for search.

Citadel Search - Feature Overview

- **Access control per documents, which otherwise would need to be implemented (or pay a commercial ES plugin).**
  - Via instance configuration and document content.

- **File content and metadata extraction via Apache Tika**
  - https://tika.apache.org/2.1.0/formats.html

- **Invenio based, and therefore customizations are easy to integrate.**
  - For example, change the authentication method from CERN SSO to other based on OAuth or many others supported by Invenio.

Citadel Search - Feature Overview

- **Ease of deploying**
  - Docker and Compose files
  - OKD templates
  - Helm charts
  - https://cern-search-admin.docs.cern.ch
  - https://cern-search.docs.cern.ch
- **Support**
  - Used at CERN which means maintenance is critical and applicable to external instances. Same for the base framework (Invenio).
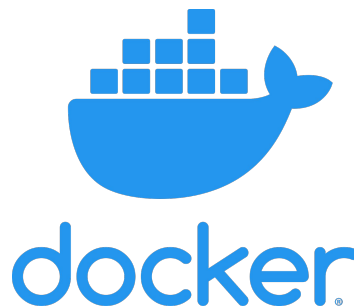
Methodology

Citadel customizations

- Elasticsearch mappings for indexing
- JSONSchema mappings for validation
- Filters
- Aggregations
- Sorting
- Custom deployment
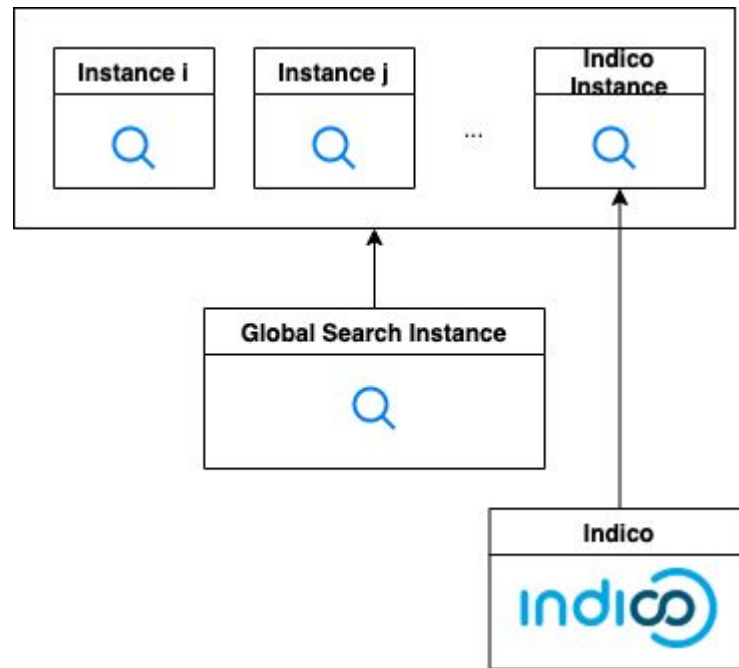
The collaboration

- **Fermilab**
- **BNL**
- **CERN - Indico**
- **CERN - Search as a service**

Methodology

Livesync: search engine population plugin

- An abstraction composed of multiple agents that may populate different targets.

Citadel plugin: Elastic-search specific backend

- Feeds livesync records into Elasticsearch specialized metadata.
- Responsible for the notion of indexable and non-indexable attributes.
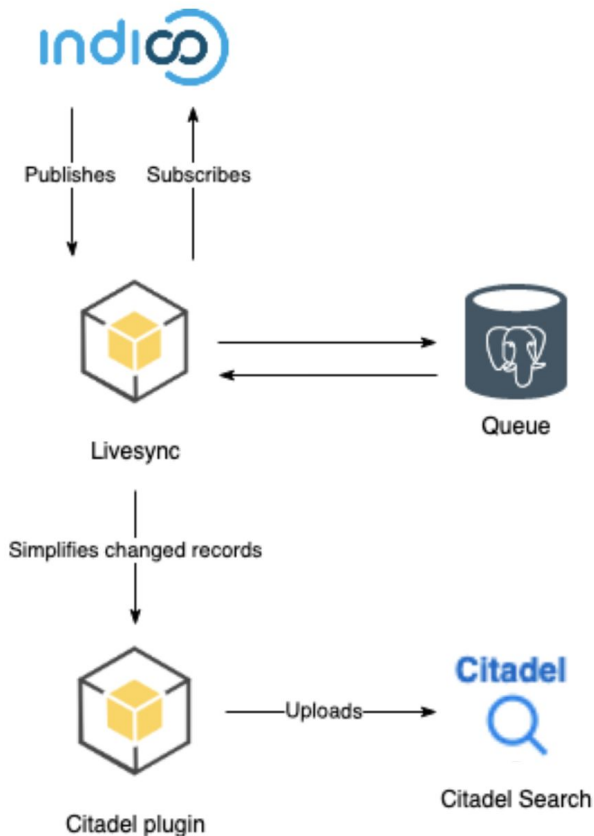
Methodology

Keeping the data fresh

- Change-tracking is enabled by attaching a set of listeners to Indico.

- The synchronization starts periodically and simplifies the queue of records before uploading them into the existing search engine.
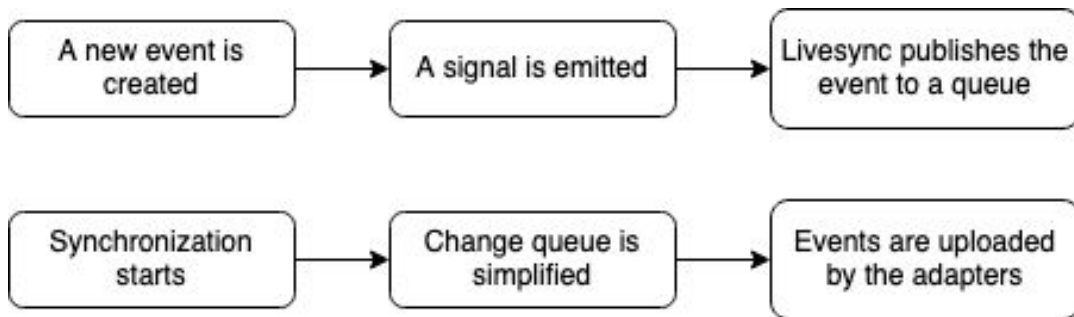
Mappings

Indico is composed of multiple search relevant entries, being:

- Events
- Contributions
- Materials (files associated with a contribution or event)
- Notes (description of meeting entries, also known as *minutes*)

```
event_title: My event title
event_description: My description
location: My venue
chairpersons: John, Peter, Maria
```

```
contribution_title: My contribution title
contribution_description: My description
location: My venue
authors: Paul, Joe
```

Result

Aggregations

- Groups of aggregated content and ranges with an associated filter, transparent to the user query-string.

Filters/Placeholders

- Arbitrary attribute matching.

Materials

- Text-based search on any Indico extracted file content.

Role-based control

- Limit the access to the user access control list (ACL) and avoid exposing restricted information.

Affiliation

- [ ] Desy (1)
- [ ] Gde (1)
- [ ] Slac (1)
- [ ] University of california,santa cru...

Category

- [ ] Home (29)
- [ ] Physics and Detectors (16)
- [ ] Detector R&D collaborations (15)
- [ ] Conference and Workshops (7)
- [ ] JSPS - Detector R&D towards th...
- [ ] TPC R&D (6)
- [ ] Conferences (3)
- [ ] Conferences, Workshops and Ev...
- [ ] 2017 (2)
- [ ] ILC-related Presentations Given ...

Person

- [ ] Bruce Andrew Schumm (1)
- [ ] Bruce Schumm (1)
- [ ] Dr Barry Barish (1)
- [ ] Felix Sefkow (1)
- [ ] Michael Peskin (1)

title:LHC type:conference

title	The title an event, contribution, etc.)
person	A speaker, author or event chair's name
affiliation	A speaker, author or event chair's affiliation
type	An entry type (such as conference, meeting, file, etc.)
venue	Name of the venue
room	Name of the room
address	Address of the venue
file	Name of the attached file
keyword	A keyword associated with an event
category	The category of an event

LCG SC4 / Initial Service Preparation Workshop - Asia-Pacific
Workshop to review SC3 experience and prepare for SC4 and Initial LHC Computing Service
📅 30 November 2005 08:00
🗂 Home » Conferences, Workshops and Events » Conferences

Où est la matière noire?
théories de matière noire, revues à l'éclairage des recherches directes récentes et des résultats du LHC
📅 08 November 2013 12:30
🗂 Home » Institutions » DAPNIA

The LHC Early Phase for the ILC
The LHC will produce its first physics data at 14 TeV in 2008 and has the promise of making groundbreaking It is now time to examine the scenarios that may occur in the early phase of the LHC operation and investigate
📅 12 April 2007 - 14 April 2007
🗂 Home » Conference and Workshops » 2007

LCTPC-pixel meeting #21
Thursday 28 January 2016 West Coast East Coast W.Europe Japan 05.00 08.00 14.00 22.00 PLEASE NOTE: The LCTPC Pixel meetings are done by Vidyo in my room: Room number: 1010700723 Link: h
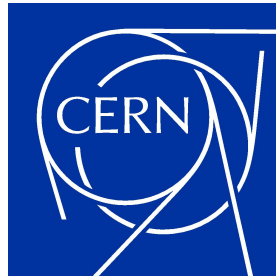📅 28 January 2016 13:00

Future Work

- **OpenSearch**

- **Documentation for non-CERN ecosystem deployment and less tech savvy users**

Thank you

Questions?

pedro.lourenco@cern.ch **(Indico)**

carina.antunes@cern.ch **(Citadel Search)**

home.cern