Digital library definitions

- Collections are stored in **digital formats**: NO prints, NO microfilms, NO tapes...

- A type of **Information Retrieval** system

- A **Virtual Organisation** with Targeted communities

- **Repository types**
  - **Institutional** Document repositories
  - World-wide **subject**-based information systems

- Many technical options: local, on the cloud, commercial, open source, etc
Different concepts

Content types: **Born digital** versus **Converted** into digital

Archive concept: digital **Library** versus digital **Archive**

Open access: **Green** versus **Gold**

- Document Repositories manage eprints
- Library Systems manage series, books, journals
- Multimedia Systems for photos and videos (MAM)
- Document Management System (GED)
- Data Research repositories

→ **Hybrid** systems manage both **Electronic** resources and **Traditional** print material
Different concepts

BORN DIGITAL

CONVERTED TO DIGITAL
Different concepts

DIGITAL LIBRARY

DIGITAL ARCHIVE
Different concepts

GREEN OPEN ACCESS

GOLD OPEN ACCESS
Important Standards

1. **Content representation**: MARC, DC, JSON
2. **Exchange protocols**: Z39.50; OAI-PMH between Data and Service providers
3. **Interoperability** with SWORD: Simple Web-service Offering Repository Deposit
4. **Identifiers**: ISBN, DOI, ORCID, etc
5. **Preservation** of metadata: METS with descriptive, structural and administrative content in the **OAIS** ref. Model (ISO 16363)
6. **Licensing** with Creative Commons

- <1 million records of articles, books, theses, photos, objects and more material produced at CERN
- Powered by Invenio
- Institutional ; Hybrid ; Born & not-Born digital ; Library & Archive ; Green & Gold
- http://cds.cern.ch
- http://videos.cern.ch
Ex 2: Inspire (2007 - )

- High Energy Physics information system run by CERN, DESY, FNAL, SLAC...
- Powered by Invenio, metadata curation since the 1960s (in SPIRES)
- Disciplinary; Hybrid; Born digital; Library; Green
- http://inspirehep.org
More

Digital Libraries?

- Eprints ArXiv?
- Zenodo?
- Google Books?
- YouTube?
- Internet Archive?

http://www.internetlivestats.com/
Software supporting Library Systems

Eprints
Dspace
Fedora
Greenstone
Koha
Invenio
...

1. Building, maintaining, managing, running DLs
2. Ingest, Preservation and Access for locally produced academic outputs
3. Implementing interoperability
4. Following up standards
5. High quality content: issue of supporting curation processes
6. Dissemination is organized and controlled
Why CERN?

A natural place to start with!

- 1954: Laboratory birth
- 1989: invention of the World Wide Web
- 1991: SPIRES, first database on the Web
- 1993: CERN Preprint Server birth
- 1996-2000: addition of Books, Periodicals, internal Notes and Multimedia to CDS
- 2002: CDSware SW released open source
- 2006: CDSware becomes Invenio, international collaboration
- 2013: Tind Spin-of sales Invenio-based services
Features developed for CERN

- **Scalable** search engine: multi-million records; metadata & full text queries
- **Flexible** metadata representation (MARC or JSON - native)
- **Collaborative** features & Internationalization
- **Books** management and circulation (v1.x)
- **Open Source**, MIT license (v3), open to customization with RESTful APIs
- **Hybrid**: eprint repo + library system + multimedia server + doc mgmt
The Usual Workflows
Global flows of a Library System

- **Author**
- **Sources**
- **Ingestion**
- **Librarian**
- **Processing**
- **DB System**
- **Curation**
- **Dissemination**
- **User**

The main components
Ingestion overview

Web Submission Interfaces workflows and functions

1. **Batch Import**
2. **Web Submission**
   - **Convert**
   - **Standardize**
   - **Schedule**
   - **Upload**

- **Non OAI Source**
- **OAI Source**
- **Author**

**Session Access**

**Metadata**

**Data Files**
Processing overview

Data Files
- Classification
- Extraction

Metadata
- Community
- Formats
- Indexes
- Ranks

Clusters

Indexes
- logical fields
- ranking
Basics of Ranking

PageRank-like and hot trends

‘most cited’ documents

inspirehep.net (500 random points)
Basics of Ranking

inspirehep.net (500 random points)
Basics of Ranking

PageRank-like and hot trends
Dissemination overview

Multi-level queries

Clusters, Indexes, etc.

Meta data

Data Files

Search UI

Tags

Baskets

Discussion

Newsletter

Circulation

User

Push & Pull

Alerts

Help pages

Statistics

Systems

OAI

RSS
From a user query to the result page
From a user query to the result page

- User query: "Rafiki AND Kahiu IN Films"
- Collection universe
- Search results page: "Records formats"
- Word indexes: forward & reverse
From a user query to the result page

User query to the result page:

1. **User** submits a query (e.g., Rafiki AND Kahiu IN Films).
2. **Query Parser** processes the query and sends it to the collection clusterer.
3. **Collection Clusterer** receives the query and sends it to the boolean intersector.
4. **Boolean Intersector** performs set operations (AND) on the collections and sends the results to the word searcher.
5. **Word Searcher** retrieves relevant words (Rafiki, Kahiu) and sends them back to the boolean intersector.
6. **Deserialize** processes the search results and returns them to the search results page.
7. **Search Results Page** displays the formatted records (Rafiki AND Kahiu IN Films).

**Word Indexes:**

- Forward & Reverse

**Record Formats:**

- Collection Universe
- Films Collection
- Search Clusterer
- Boolean Intersector
- Deserializer
- Search Results Page
From a user query to the result page

User query to the result page involves the following steps:

1. User inputs a query: "Rafiki AND Kahiu IN Films".
2. The query is parsed by a query parser.
3. The parsed query is then passed to a word searcher for Rafiki and Kahiu.
4. Each word is processed through word indexes: forward & reverse.
5. Films collection is used as input to the boolean intersector.
6. The boolean intersector combines the queries using AND, resulting in Rafiki hitsets and Kahiu hitsets.
7. The hitsets are deserialized to generate final hitsets.
8. The final hitsets are clustered by the collection clusterer.
9. The clustered results are sorted by the record sorter.
10. The sorted results are formatted by the output formatter.
11. The formatted results are presented on the search results page.
Curation overview

- Enriching
- Standardizing
- Checking records
- Extracting from files

Librarian

- Check
- Match
- Merge
- Upload
- Circulate

Edit

- Classify
- Extract
- Manage file

Data Files

- Authority
- Holdingpen

Meta data

- Knowledge Bases
- Tasks

Manage file

Upload
Access control

**Authentication management**
To support an external authentication method in addition to local accounts.

User/password checking and import of user details: group memberships, phone number, affiliation, etc.

Many authentication methods: Oauth, SSO, Shibboleth, LDAP, etc.

**Authorization management**
To manage who can do what

With Role Based Access Control (RBAC):
permissions are granted to roles, assigned to users or groups

With Firewall Like Role (FireRole):
standard ‘language’ to define permissions;
allow email /.*cern.ch/,/.*@slac.stanford.edu/
deny group badguys
Conclusion: technology change
Technology Evolution (I)

2000’s (Invenio v1) — 2010’s (Invenio v3)

<table>
<thead>
<tr>
<th>2000’s</th>
<th>2010’s</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Home made templating</strong></td>
<td><strong>User interface templating</strong></td>
</tr>
<tr>
<td><strong>XML API</strong></td>
<td><strong>RESTful API</strong></td>
</tr>
<tr>
<td><strong>MARC 21 Format for Bibliographic Data</strong></td>
<td><strong>abstract record model</strong></td>
</tr>
<tr>
<td><strong>RBAC engine</strong></td>
<td><strong>Authorization</strong></td>
</tr>
<tr>
<td><strong>OAuth2.0 Flask authentication</strong></td>
<td><strong>Bootstrap Jinja</strong></td>
</tr>
<tr>
<td>Technology Evolution (II)</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td></td>
</tr>
<tr>
<td><strong>2000’s (Invenio v1)</strong></td>
<td></td>
</tr>
<tr>
<td><strong>2010’s (Invenio v3)</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BibSched</th>
<th>Internal scheduler</th>
<th>Task manager</th>
<th>Celery</th>
<th>redis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Solr &amp; home made metadata indexes</td>
<td></td>
<td>Indexing</td>
<td>elasticsearch</td>
<td></td>
</tr>
<tr>
<td>Python with MySQL</td>
<td></td>
<td>Framework</td>
<td>Flask</td>
<td>Python with PostgresQL</td>
</tr>
</tbody>
</table>
QUESTIONS?