

# Insight into Digital Library Systems



© Udayan Rao Pawar

By Jean-Yves Le Meur  
project leader of CERN Digital Memory

# 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

INTEGRATED  
LIBRARY  
SYSTEMS

EPRINTS  
REPOSITORIES

MULTIMEDIA  
SYSTEMS

HYBRIDS  
SYSTEMS

DOCUMENT  
MANAGEMENT  
SYSTEMS

RESEARCH DATA  
SYSTEMS

# Important Standards



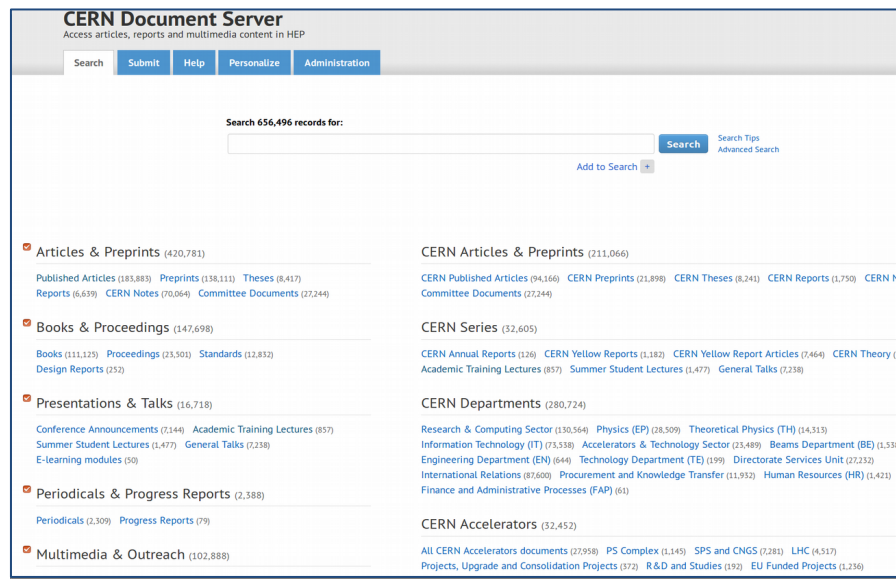
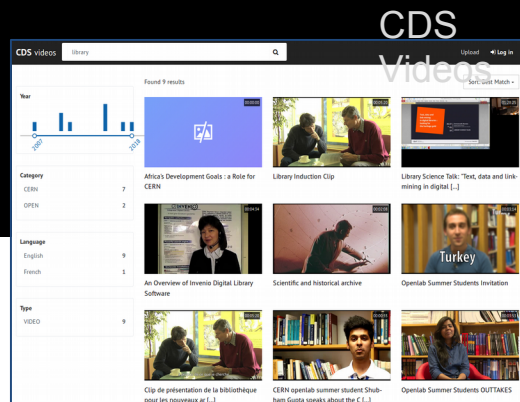
## XML-MARC: tag 100

```
<datafield tag="100" ind1=" " ind2=" " >
  <subfield code="a">Clerbaux, Barbara</subfield>
  <subfield code="e">ed.</subfield>
  <subfield code="i">INSPIRE-00314890</subfield>
  <subfield code="u">Brussels U.</subfield>
</datafield>
```

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

# Ex 1: CERN Document Server (1993 - )

- <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>

## Citations

## Author pages

standard model
[find "Phys Rev Lett 109":](#)
devenaire

CiteSummary

Dixon, Lance Jenkins

Profile Name

[View Profile](#)
[Manage Profile](#)
[Manage Publications](#)
[Add](#)

Personal Details (preferences)

Name
Lance J. Dixon

Email
lance@dixon.science.edu

Links
[http://www.science.edu/~ljd/](#)

Labels
HEP-Ph  
HEP-TH

Identifiers
BKL JID011  
INSPIRE
INSPIRE: 0076732  
ORCID: 0000-0001-0001-7103

Period
Rank
Institution

1992
50000
SLAC

1993
50000
SLAC

1997 - 1999
30000
JANOR
Princeton U.

1998 - 1997
100
SLAC

1998 - 1987
PHD
Princeton U.

1978 - 1982
US
Caltech

[Update Details](#)

Publications

Publications

External

1. Analytic calculation of Energy-Energy Correlation in  $e^+e^-$  annihilation at NLO

2. Two-Quark Production Integral in  $AB$  Correlators

3. Analytical Calculation of Energy-Energy Correlation at Next-to-Leading Order in QCD

4. The Energy of Resonant Transparencies and the Four-Loop Collinear Anomalous Dimension

5. Gauge-Like Feynman Diagrams Into Fishes

6. Two-Loop Renormalization of Quantum Gravity Simplified

7. Resurgence from the Spectrum-Counting Bootstrap

8. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

9. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

10. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

[Click here to see all](#)

standard model
[find "Phys Rev Lett 109":](#)
devenaire

CiteSummary

Dixon, Lance Jenkins

Profile Name

[View Profile](#)
[Manage Profile](#)
[Manage Publications](#)
[Add](#)

Personal Details (preferences)

Name
Lance J. Dixon

Email
lance@dixon.science.edu

Links
[http://www.science.edu/~ljd/](#)

Labels
HEP-Ph  
HEP-TH

Identifiers
BKL JID011  
INSPIRE
INSPIRE: 0076732  
ORCID: 0000-0001-0001-7103

Period
Rank
Institution

1992
50000
SLAC

1993
50000
SLAC

1997 - 1999
30000
JANOR
Princeton U.

1998 - 1997
100
SLAC

1998 - 1987
PHD
Princeton U.

1978 - 1982
US
Caltech

[Update Details](#)

Publications

Publications

External

1. Analytic calculation of Energy-Energy Correlation in  $e^+e^-$  annihilation at NLO

2. Two-Quark Production Integral in  $AB$  Correlators

3. Analytical Calculation of Energy-Energy Correlation at Next-to-Leading Order in QCD

4. The Energy of Resonant Transparencies and the Four-Loop Collinear Anomalous Dimension

5. Gauge-Like Feynman Diagrams Into Fishes

6. Two-Loop Renormalization of Quantum Gravity Simplified

7. Resurgence from the Spectrum-Counting Bootstrap

8. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

9. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

10. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

[Click here to see all](#)

standard model
[find "Phys Rev Lett 109":](#)
devenaire

CiteSummary

Dixon, Lance Jenkins

Profile Name

[View Profile](#)
[Manage Profile](#)
[Manage Publications](#)
[Add](#)

Personal Details (preferences)

Name
Lance J. Dixon

Email
lance@dixon.science.edu

Links
[http://www.science.edu/~ljd/](#)

Labels
HEP-Ph  
HEP-TH

Identifiers
BKL JID011  
INSPIRE
INSPIRE: 0076732  
ORCID: 0000-0001-0001-7103

Period
Rank
Institution

1992
50000
SLAC

1993
50000
SLAC

1997 - 1999
30000
JANOR
Princeton U.

1998 - 1997
100
SLAC

1998 - 1987
PHD
Princeton U.

1978 - 1982
US
Caltech

[Update Details](#)

Publications

Publications

External

1. Analytic calculation of Energy-Energy Correlation in  $e^+e^-$  annihilation at NLO

2. Two-Quark Production Integral in  $AB$  Correlators

3. Analytical Calculation of Energy-Energy Correlation at Next-to-Leading Order in QCD

4. The Energy of Resonant Transparencies and the Four-Loop Collinear Anomalous Dimension

5. Gauge-Like Feynman Diagrams Into Fishes

6. Two-Loop Renormalization of Quantum Gravity Simplified

7. Resurgence from the Spectrum-Counting Bootstrap

8. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

9. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

10. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

[Click here to see all](#)

standard model
[find "Phys Rev Lett 109":](#)
devenaire

CiteSummary

Dixon, Lance Jenkins

Profile Name

[View Profile](#)
[Manage Profile](#)
[Manage Publications](#)
[Add](#)

Personal Details (preferences)

Name
Lance J. Dixon

Email
lance@dixon.science.edu

Links
[http://www.science.edu/~ljd/](#)

Labels
HEP-Ph  
HEP-TH

Identifiers
BKL JID011  
INSPIRE
INSPIRE: 0076732  
ORCID: 0000-0001-0001-7103

Period
Rank
Institution

1992
50000
SLAC

1993
50000
SLAC

1997 - 1999
30000
JANOR
Princeton U.

1998 - 1997
100
SLAC

1998 - 1987
PHD
Princeton U.

1978 - 1982
US
Caltech

[Update Details](#)

Publications

Publications

External

1. Analytic calculation of Energy-Energy Correlation in  $e^+e^-$  annihilation at NLO

2. Two-Quark Production Integral in  $AB$  Correlators

3. Analytical Calculation of Energy-Energy Correlation at Next-to-Leading Order in QCD

4. The Energy of Resonant Transparencies and the Four-Loop Collinear Anomalous Dimension

5. Gauge-Like Feynman Diagrams Into Fishes

6. Two-Loop Renormalization of Quantum Gravity Simplified

7. Resurgence from the Spectrum-Counting Bootstrap

8. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

9. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

10. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

[Click here to see all](#)

standard model
[find "Phys Rev Lett 109":](#)
devenaire

CiteSummary

Dixon, Lance Jenkins

Profile Name

[View Profile](#)
[Manage Profile](#)
[Manage Publications](#)
[Add](#)

Personal Details (preferences)

Name
Lance J. Dixon

Email
lance@dixon.science.edu

Links
[http://www.science.edu/~ljd/](#)

Labels
HEP-Ph  
HEP-TH

Identifiers
BKL JID011  
INSPIRE
INSPIRE: 0076732  
ORCID: 0000-0001-0001-7103

Period
Rank
Institution

1992
50000
SLAC

1993
50000
SLAC

1997 - 1999
30000
JANOR
Princeton U.

1998 - 1997
100
SLAC

1998 - 1987
PHD
Princeton U.

1978 - 1982
US
Caltech

[Update Details](#)

Publications

Publications

External

1. Analytic calculation of Energy-Energy Correlation in  $e^+e^-$  annihilation at NLO

2. Two-Quark Production Integral in  $AB$  Correlators

3. Analytical Calculation of Energy-Energy Correlation at Next-to-Leading Order in QCD

4. The Energy of Resonant Transparencies and the Four-Loop Collinear Anomalous Dimension


5. Gauge-Like Feynman Diagrams Into Fishes

6. Two-Loop Renormalization of Quantum Gravity Simplified

7. Resurgence from the Spectrum-Counting Bootstrap

8. Resurgence of the  $g$  and  $1/g$  in  $AB$  non-perturbative

9. Resurgence of the  $g$  and  $1/g$  in  $AB</$



Welcome to **INSPIRE**

HEP

[HEPNAMES](#)
[INSTITUTIONS](#)
[CONFERENCES](#)
[JOURNALS](#)

## HEP Search

### High-Energy Physics Literature Database

Use "find" for SPIRES-style search ([other tips](#))

Brief format

Recherche

[Easy Search](#)  
[Recherche avancée](#)

find | Phys.Rev.Lett.,105" :: [davantage](#)

#### HOW TO SEARCH

SPIRES syntax is (mostly) supported (requires "find")

- find a richter, b and t quark and date > 1984
- find | phys.rev.,D50,1140 or j hep.0903,112
- find eprint arxiv:1007.5048 (Note the plots available on the detailed record)
- find fulltext:"quark-gluon plasma" (Note new "fulltext" operator)
- find a ellis and refersto a witten (Note "refersto")
- find a kane and citedby title SUSY and topcite 200+ (Note "citedby")

New techniques:

- 1985 richter quark multiplicity
- arXiv:1007.5048
- citedby:author:ellis -refersto:author:witten
- author:randall | author:sundrum cited:450->1350

Additional Help:

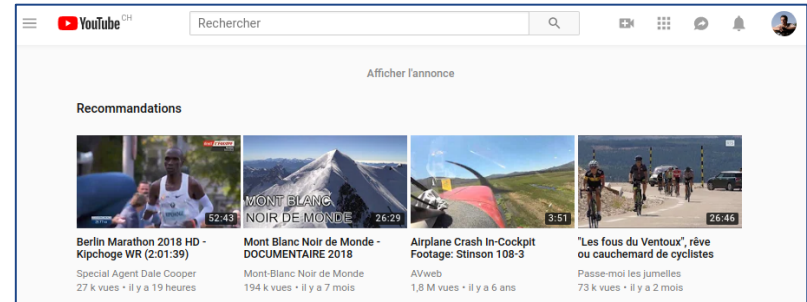
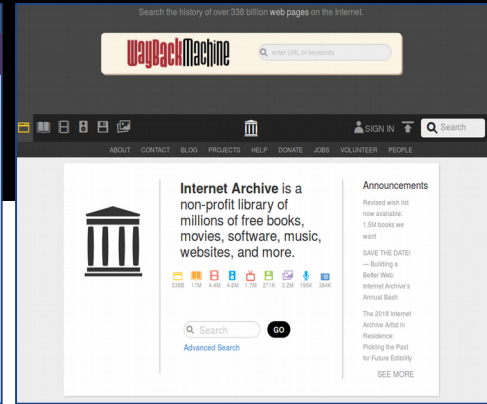
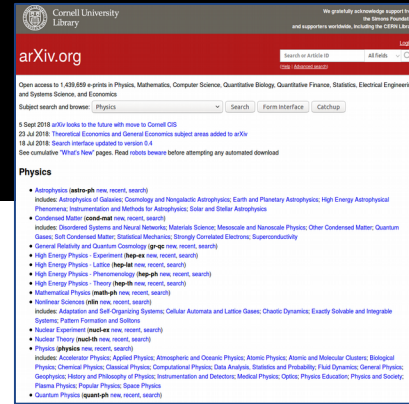
[More search tips](#) and [full help](#)

#### INSPIRE UPDATES

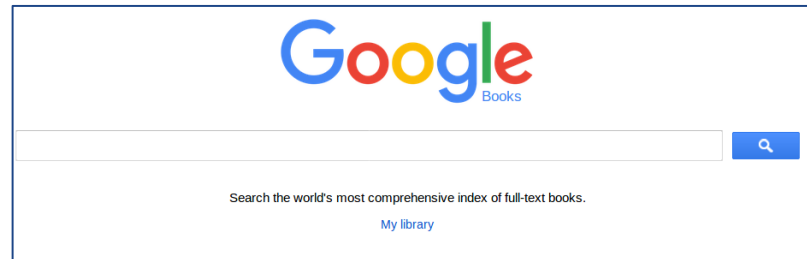
See our blog at [blog.inspirehep.net](http://blog.inspirehep.net) for updates on new features and other news. You can also follow us at [@inspirehep](https://twitter.com/inspirehep) on twitter. To send us feedback use [feedback@inspirehep.net](mailto:feedback@inspirehep.net). The data in INSPIRE is updated daily. To request corrections to data in INSPIRE, write us at [help@inspirehep.net](mailto:help@inspirehep.net). INSPIRE superseded SPIRES in 2012.

# 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-off sales Invenio-based services

# Features developed for CERN

# INVENIO

invenio-software.org  
v1 → v3

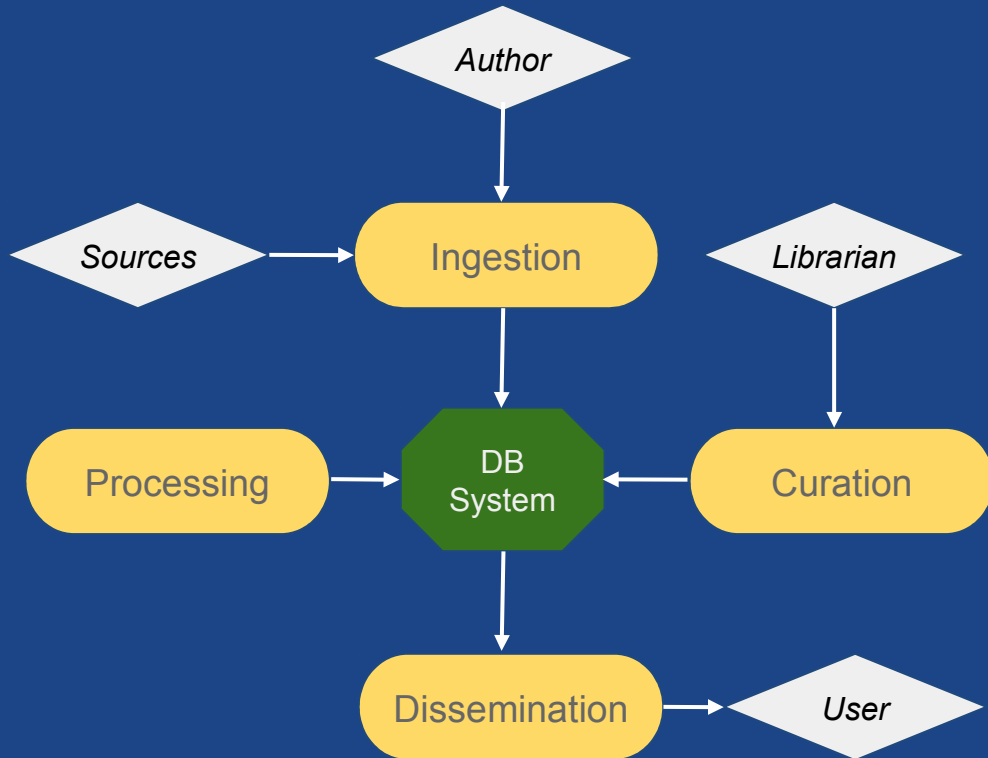


- **Scalable** search engine : multi-million records ; metadata & full text queries
- **Flexible** metadata representation (MARC or JSON - native)
- **Collaborative** features & Internationalization
- **Books** management and circulation (v1.x)
- **Open Source**, MIT license (v3), open to customization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

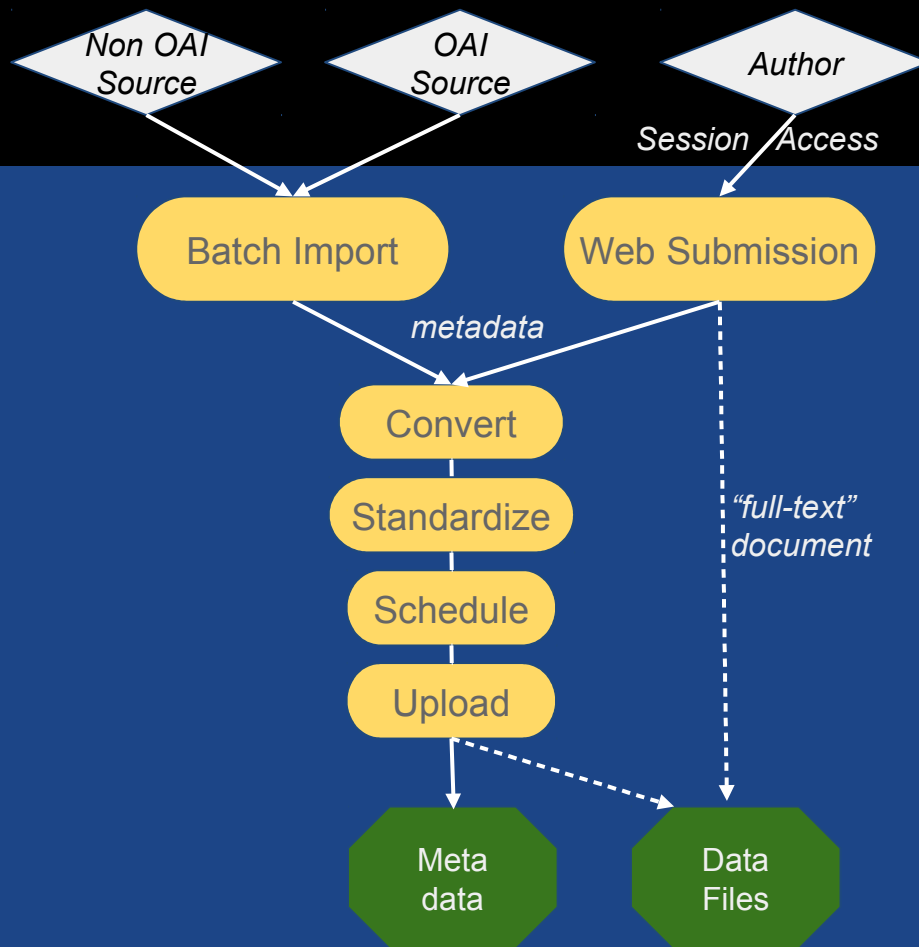
# The Usual Workflows



# Global flows of a Library System



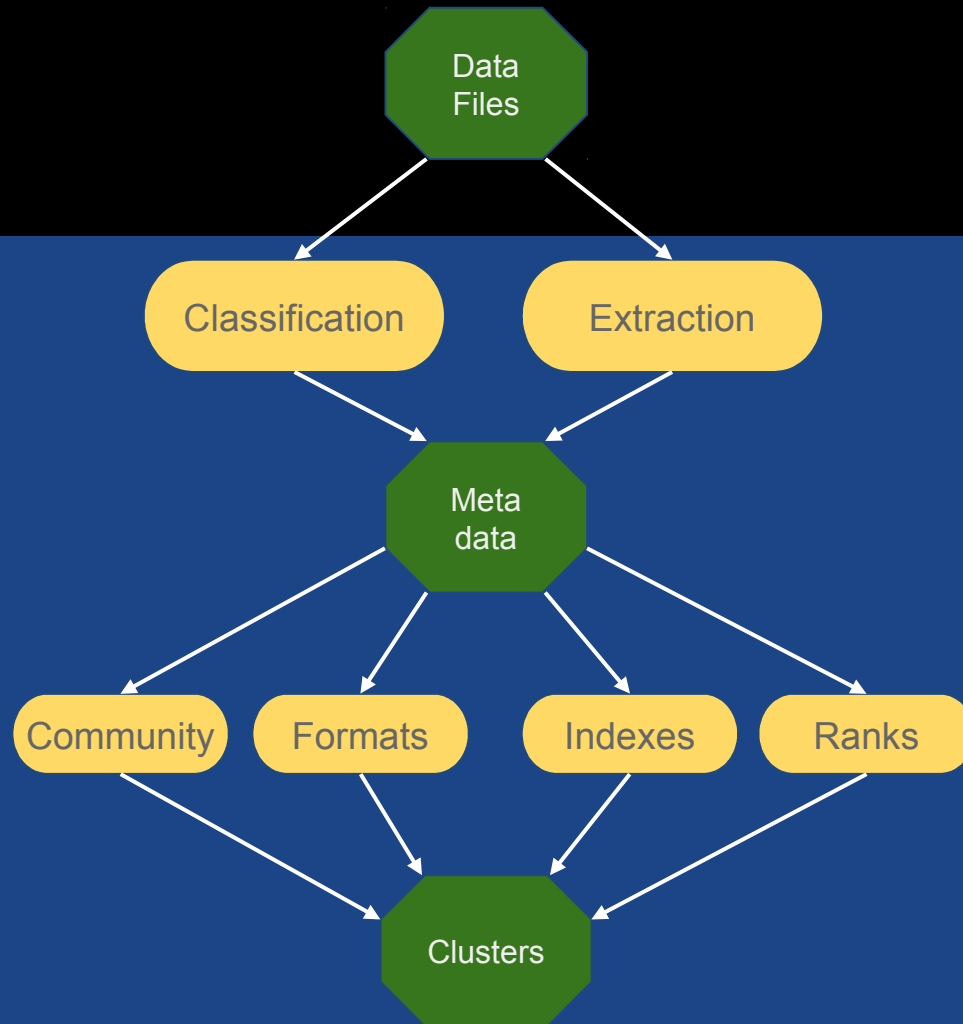
The main  
components



Ingestion  
overview

Web Submission  
Interfaces  
workflows  
and functions

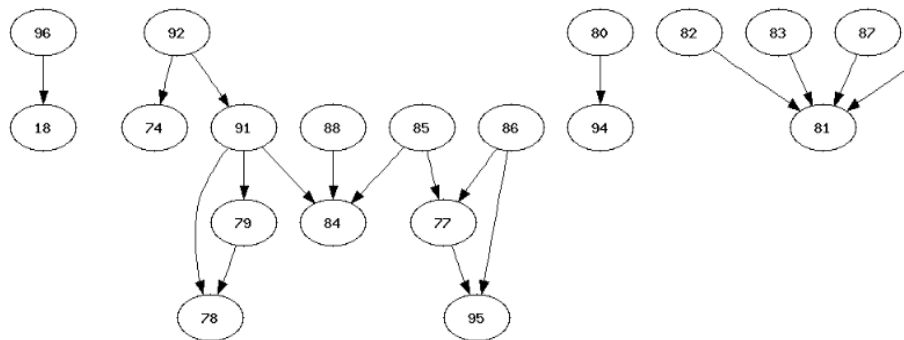
# Processing overview



Indexes  
logical fields  
ranking

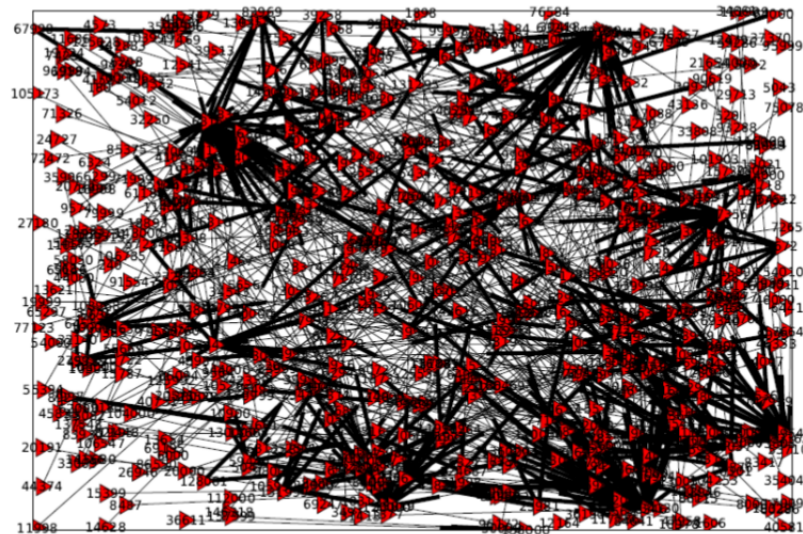
# Basics of Ranking

## PageRank-like and hot trends



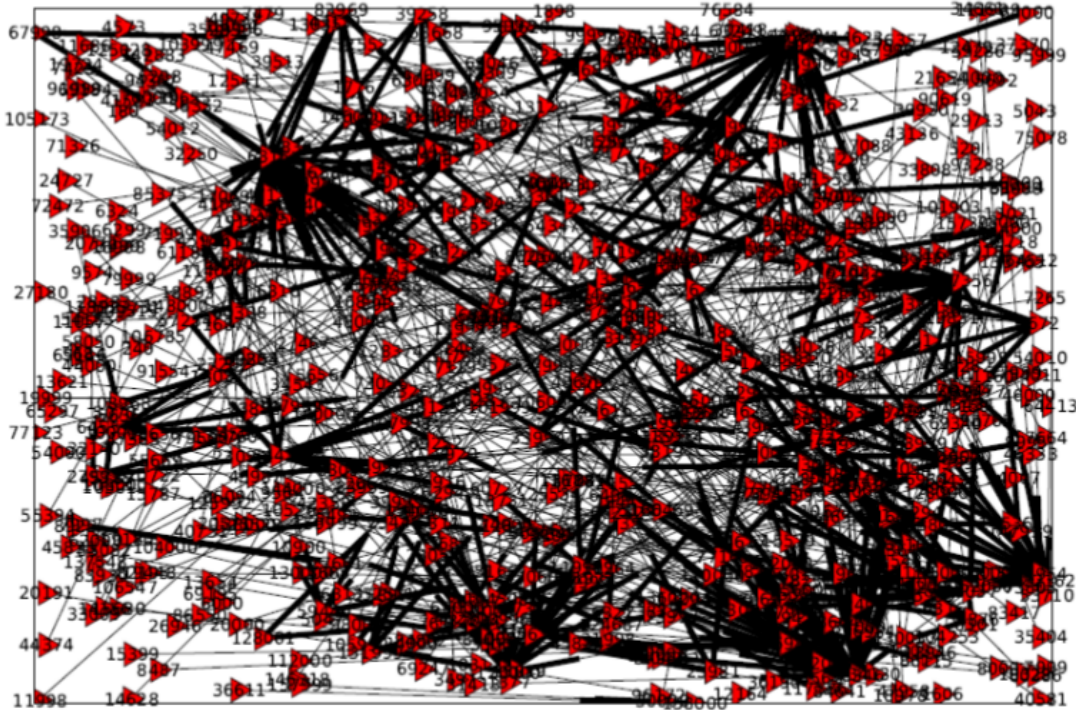
## ‘most cited’ documents

inspirehep.net (500 random points)



# Basics of Ranking

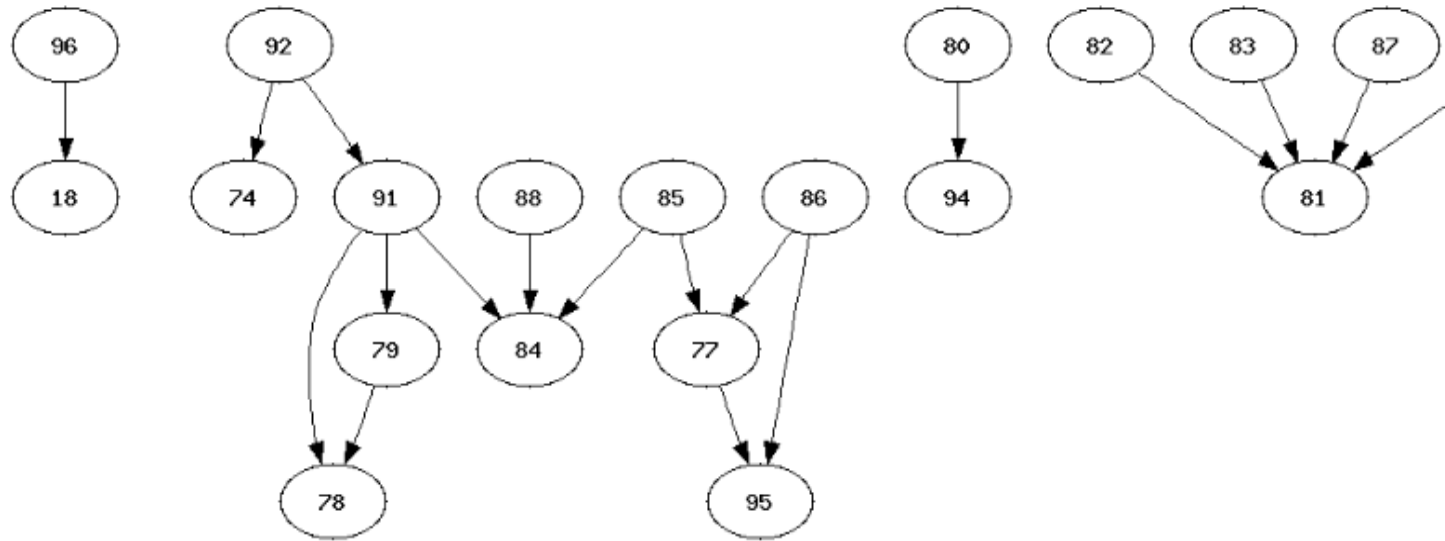
inspirehep.net (500 random points)



# Clustering Big data

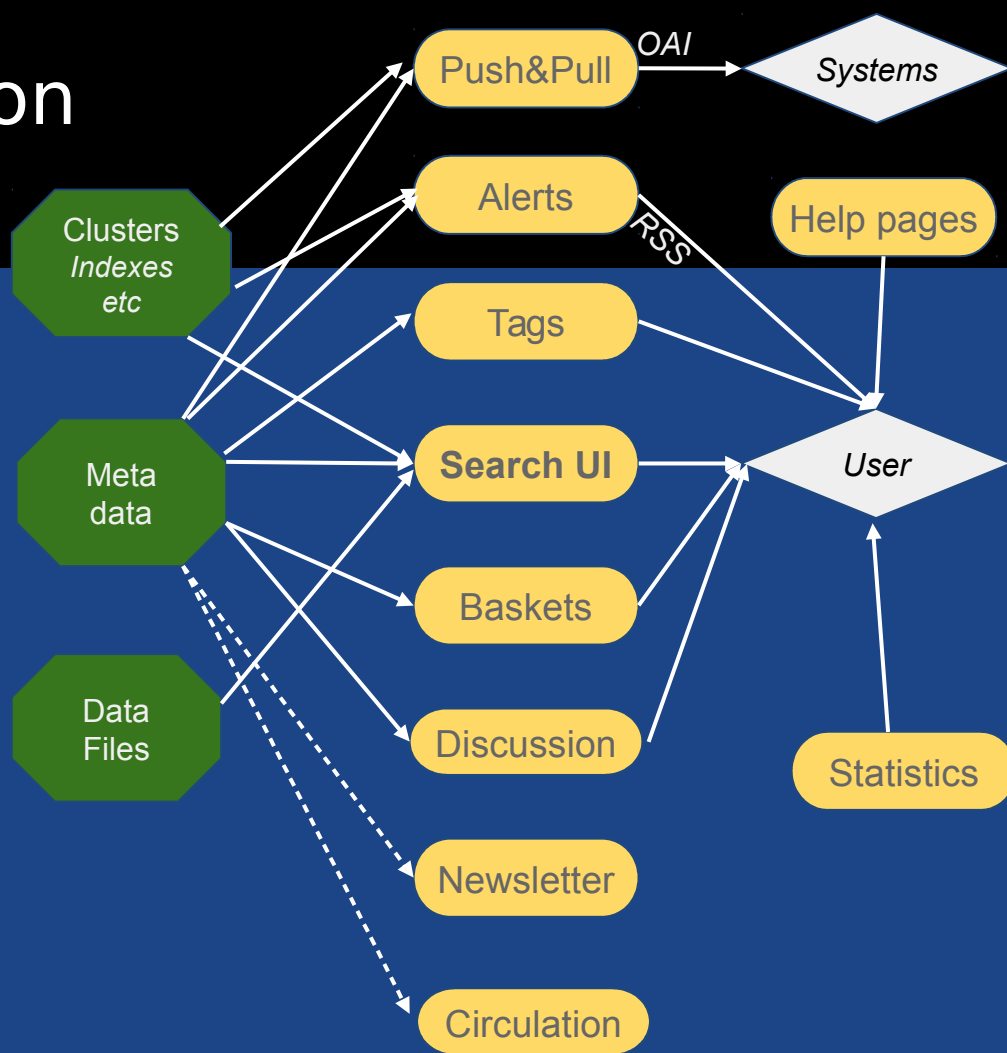
# Basics of Ranking

## PageRank-like and hot trends



# Dissemination overview

## Multi-level queries



# From a user query to the result page



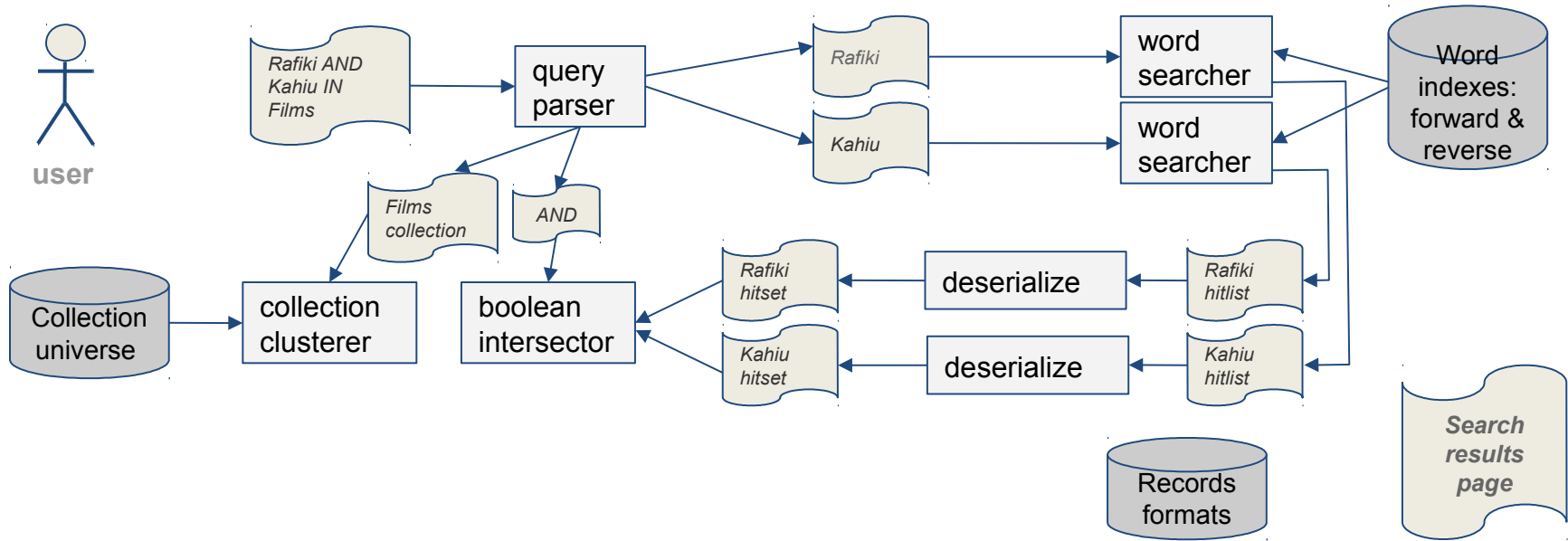
user



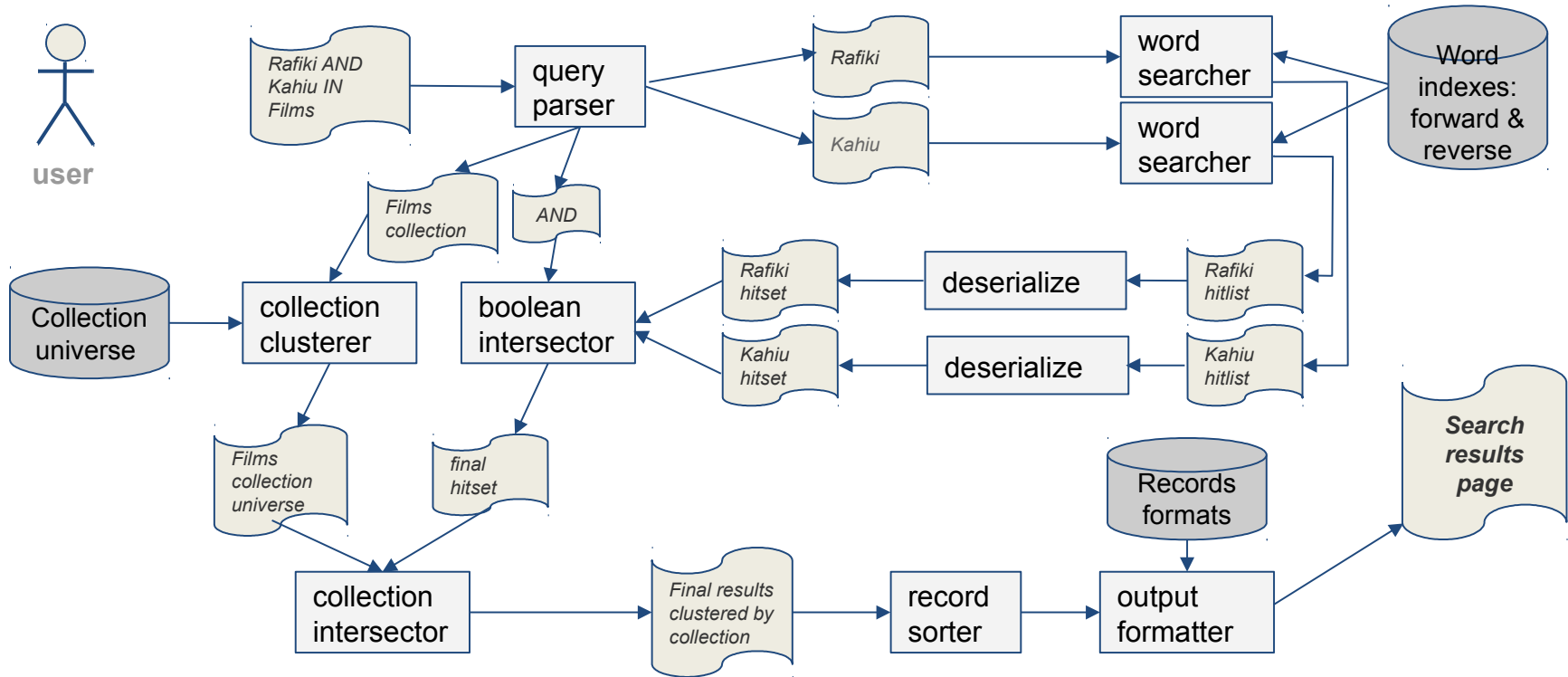
# From a user query to the result page



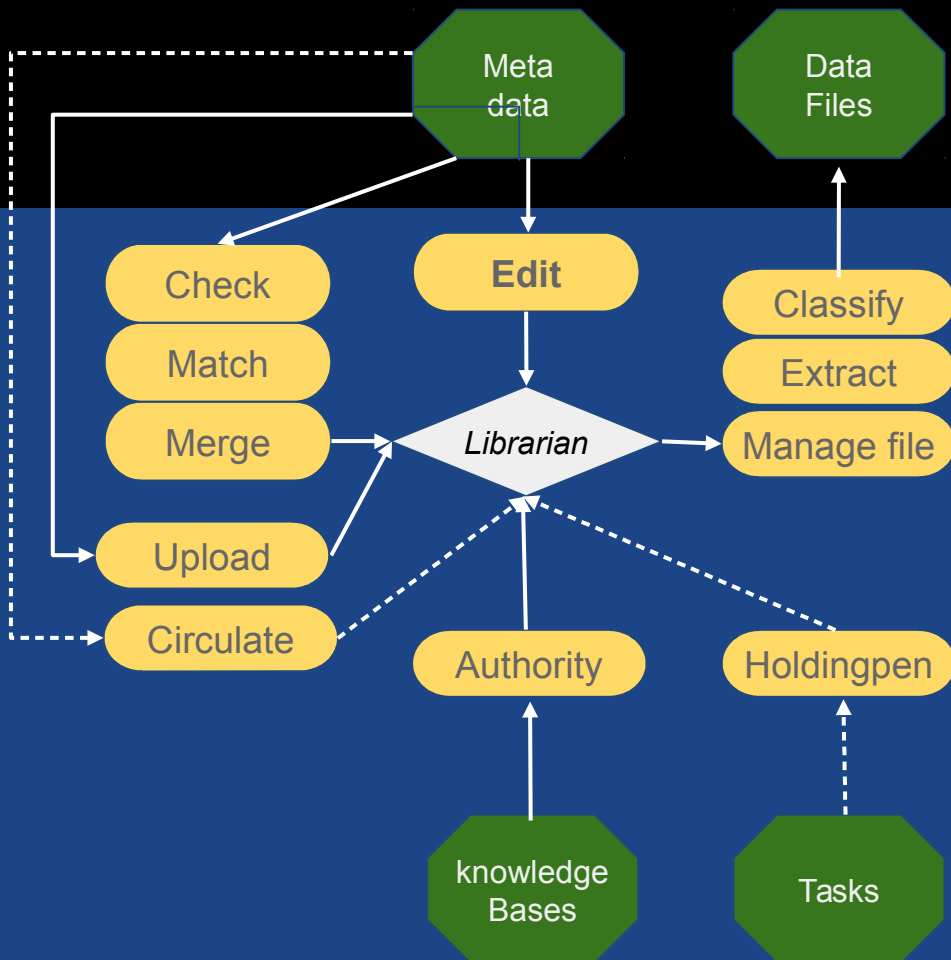
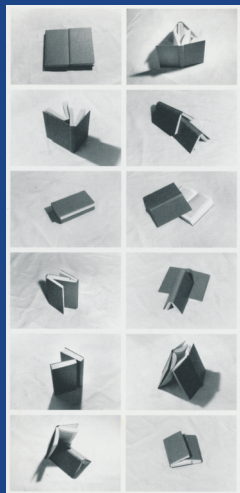
# From a user query to the result page



# From a user query to the result page



# Curation overview



Enriching  
Standardizing  
Checking records  
Extracting from files

# 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)

	Home made templating		User interface		templating	
	XML API		API			
	MARC 21 Format for BIBLIOGRAPHIC DATA Library of Congress Network Development and MARC Standards Office		Data model		abstract record model	
	 RBAC engine		Authorization		OAuth2.0	Flask authentication

# Technology Evolution (II)

2000's (Invenio v1)  2010's (Invenio v3)

 BibSched	Internal scheduler		Task manager	 Celery	 redis
 Solr	& home made metadata indexes		Indexing	 elasticsearch	
 MySQL	Python with MySQL		Framework	 Flask	Python with PostgreSQL

A close-up photograph of a lioness cub in a savanna environment. The cub is looking directly at the camera with large, orange-brown eyes. The background shows a sunset or sunrise with a warm orange glow on the horizon and a clear blue sky above. The cub's fur is a light brown color, and its whiskers are prominent. The ground is dry and sandy with some sparse vegetation.

QUESTIONS ?