Inspire

Tibor Šimko

<tibor.simko@cern.ch>

HEPiX Spring 2008 meeting, Geneva, Switzerland

May 5-9, 2008

Current HEP Information Systems

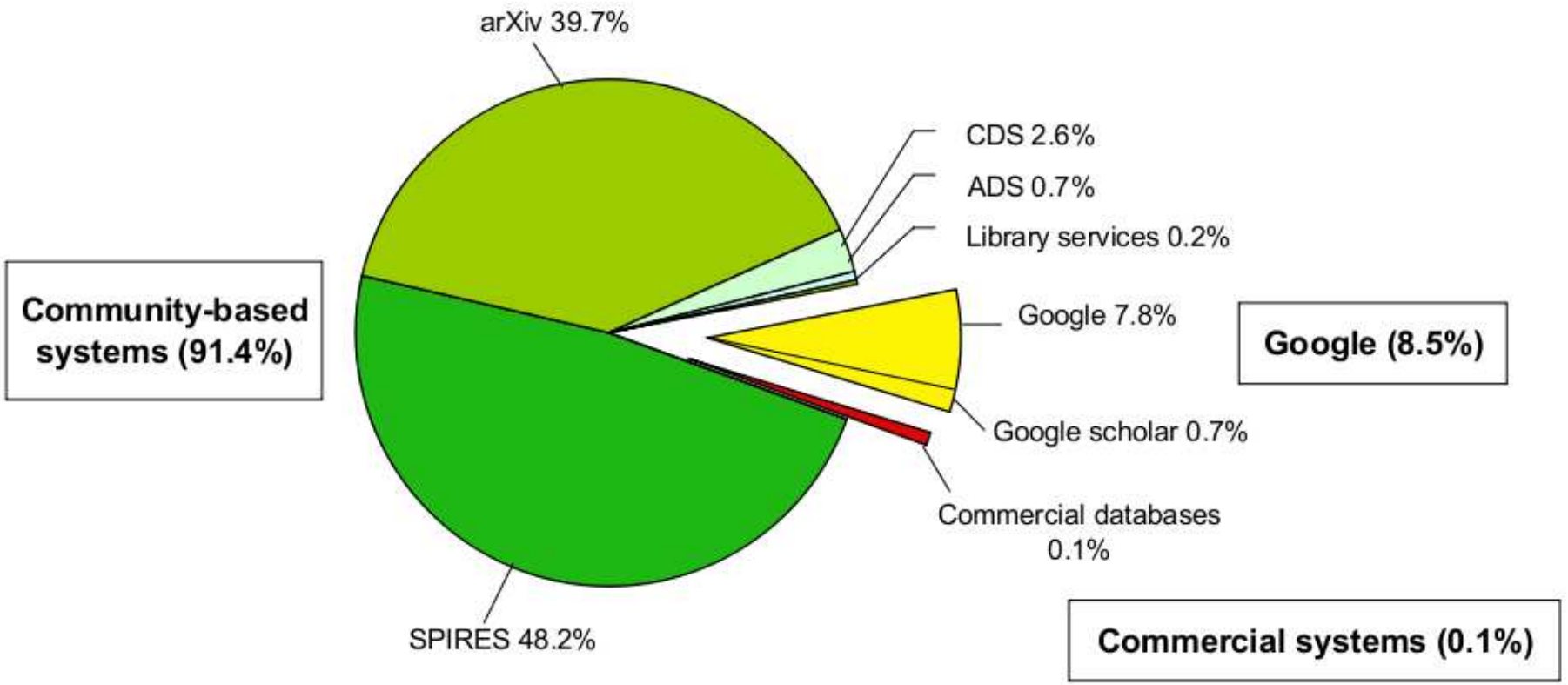
- community-based HEP scientific document servers:
  - **arXiv** (LANL, now Cornell)
  - **SPIRES** (SLAC, Fermilab, DESY)
  - **CDS** (CERN)
  - **ADS** (SAO/NASA, Harvard)
  - **KISS** (KEK)
- publisher-driven HEP scientific document servers:
  - **PROLA** (APS), etc
- generalist services:
  - **Google** Web, Google **Scholar**
- . . . end user?
- . . . synergies between community-based services?

HEP Info Systems User Poll I

- conducted concurrently by CERN, DESY, Fermilab, SLAC
- run from 30th April to 11th June 2007
- online advertizement, mailing lists
- **2,115 participants** ($\sim$10% of active HEP community)
  - ⋆ 61% theory, 22% experiment, 6% software, etc
  - ⋆ 27% US, 10% DE, 8% IT, 7% UK, 5% CERN, etc
  - ⋆ 57% everyday, 26% few times a week, 6% once a week, etc
- . . . to get to know user information habits and needs
- . . . first results just published [`arxiv:0804.2701`]

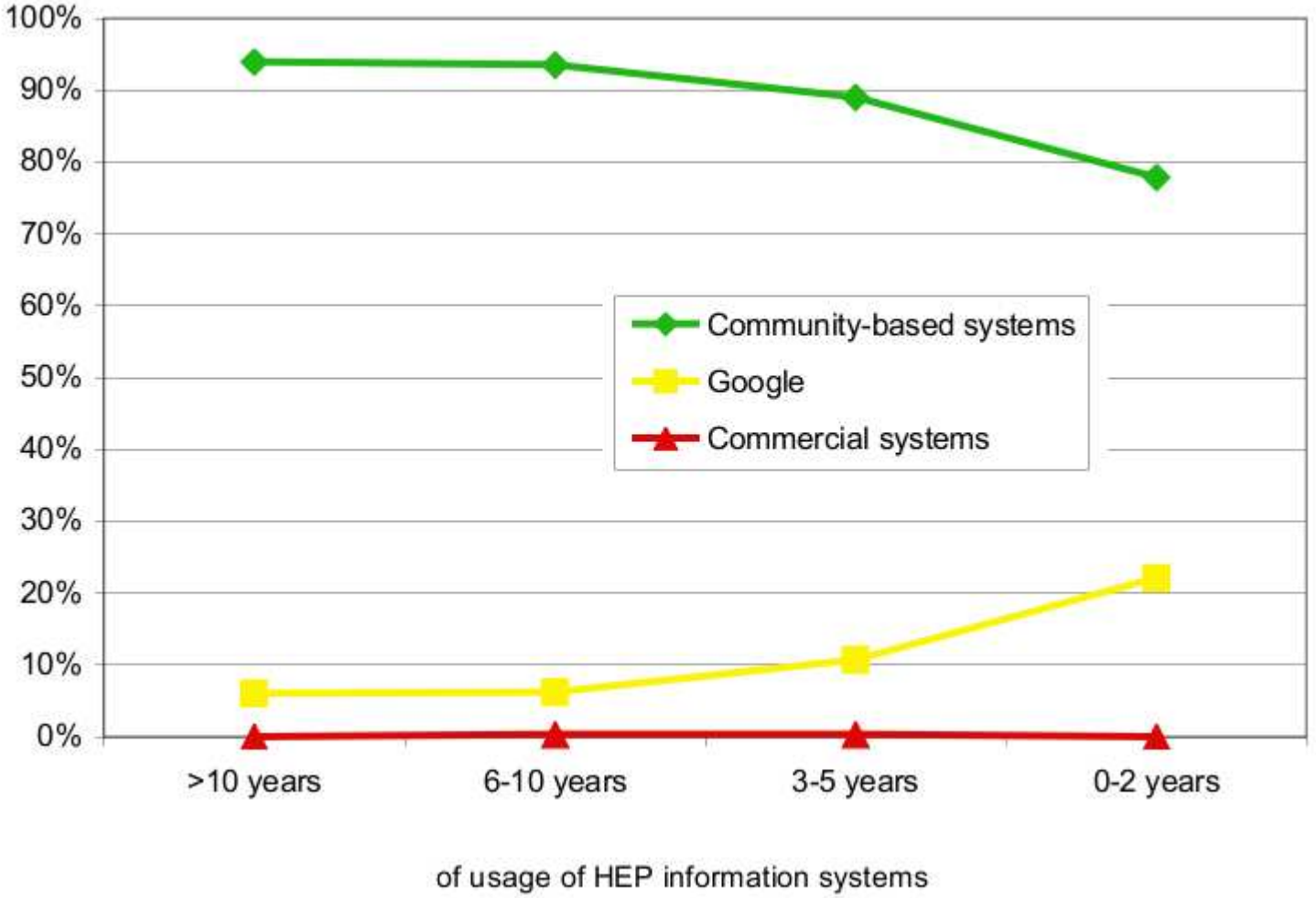
HEP Info Systems User Poll II

- E.g. what is the most used HEP information system? (“first reflex”)



HEP Info Systems User Poll III

- E.g. influence of user age on the searching habits? ("first reflex" of the "Google generation"?)



A brief history of SPIRES HEP

- **1974**: the birth of SPIRES-HEP database
  - based on SPIRES DBMS
  - IBM mainframe, command line interface
  - service run jointly by SLAC and DESY
- **1980s**: email interface to SPIRES
- **1991**: [the birth of arXiv; P. Ginsparg]
- **1991**: web interface to SPIRES
  - first US Web server
  - the “killer app” demonstrating what Web can bring (T. Berners-Lee)
- **1994**: addition of citation services
- **2000**: addition of summary formats (cite summary)

A brief history of CDS Invenio

- **1993**: CERN preprint server on the Web
- **1996**: CERN library on the Web
  - addition of books, periodicals
  - web interface on top of ALEPH DL DBMS
- **2000**: CERN Document Server
  - addition of multimedia material
  - sister web application for meeting agendas
- **2002**: first public **CDSware** release (GNU GPL)
  - Python, MySQL RDBMS
- **2004**: first public **CDS Indico** release (sister product)
- **2006**: CDSware becomes **CDS Invenio**
  - SW targeting large repositories (1M+)
  - SW used by 20+ institutions and libraries worldwide

A birth of the collaboration

- **SPIRES HEP** *(Perl, SPIRES DBMS)*
  - noise-free, high-quality metadata curation
  - human-proofed author affiliation and citation data
  - most-used system (48% “first reflex” users)
    (10+ times more searches on SPIRES than on CDS)
  - aging technology, scalability and maintenance issues
- **CDS Invenio** *(Python, MySQL RDBMS)*
  - established modern open source digital library SW
  - very good performance and scalability
  - already provides most wanted features
- **SLAC/Fermilab/DESY and CERN cataloguing**
  - institutional vs subject repositories
  - duplication of cataloguing and enrichment work

The timeline

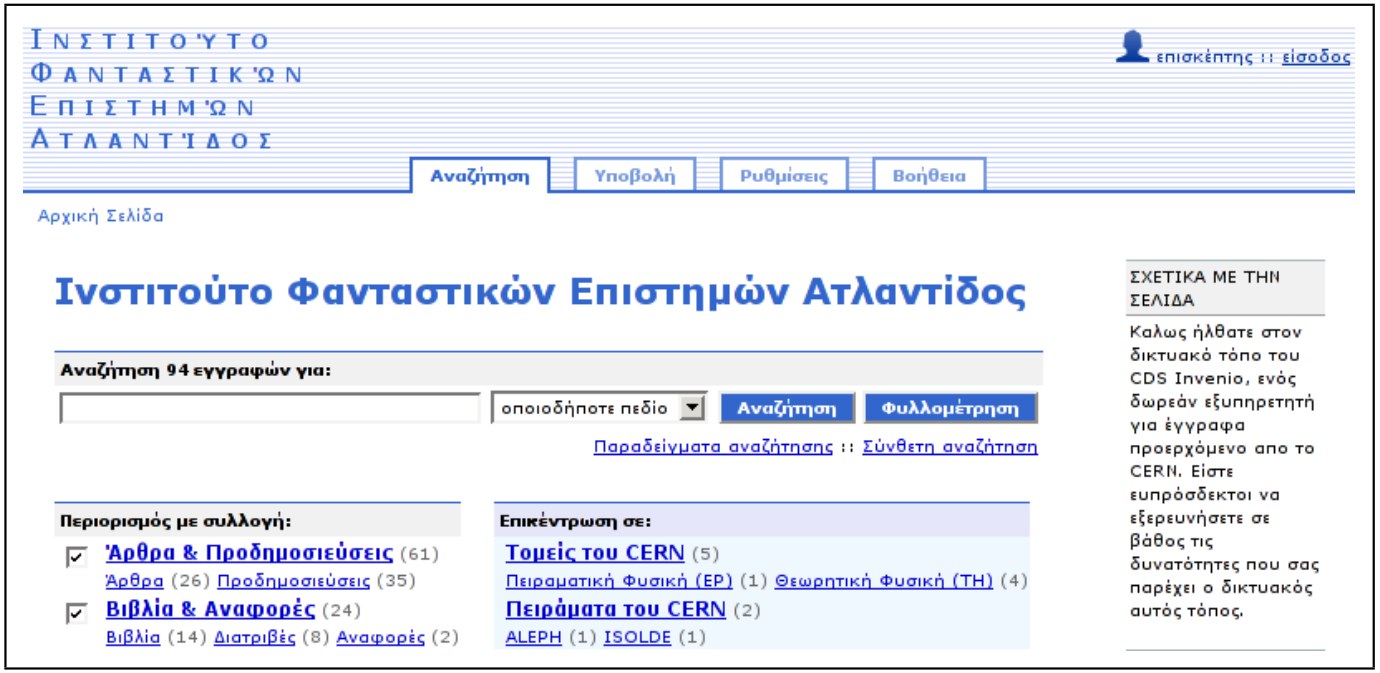
- **10-13 May 2007**: 1st *HEP/PPA Information Resource Summit* (SLAC)
  - April-June 2007: HEP Information Systems Poll
- **from June 2007**:
  - phone conferences twice a month
  - mailing lists & wiki
  - code repository
- **8-12 October 2007**: 1st Inspire workshop (CERN)
- **18-22 February 2008**: 2nd Inspire workshop (CERN)
- **20-21 May 2008**: 2nd *HEP/PPA Information Resource Summit* (DESY)
- **22-23 May 2008**: 3rd Inspire workshop (DESY)

People behind Inspire

- **SLAC**:
  - Ann Redfield, Pat Kreitz
  - Travis Brooks, Mike Sullivan
- **Fermilab**:
  - Heath O'Connell, Rob Atkinson
- **DESY**:
  - Annette Holtkamp, Zaven Akopov
- **CERN**:
  - Jens Vigen, Salvatore Mele
  - Tim Smith, Jean-Yves Le Meur, Tibor Šimko
  - Marko Niinimäki, Radoslav Ivanov
- . . . library cataloguers, curators, Invenio programmers

Inspire Phase I (summer 2007)

- *goal*: study of technical feasibility
- comparison of existing SPIRES and Invenio systems
- comparison of SLAC, Fermilab, DESY, CERN workflows
- . . . concluded positively

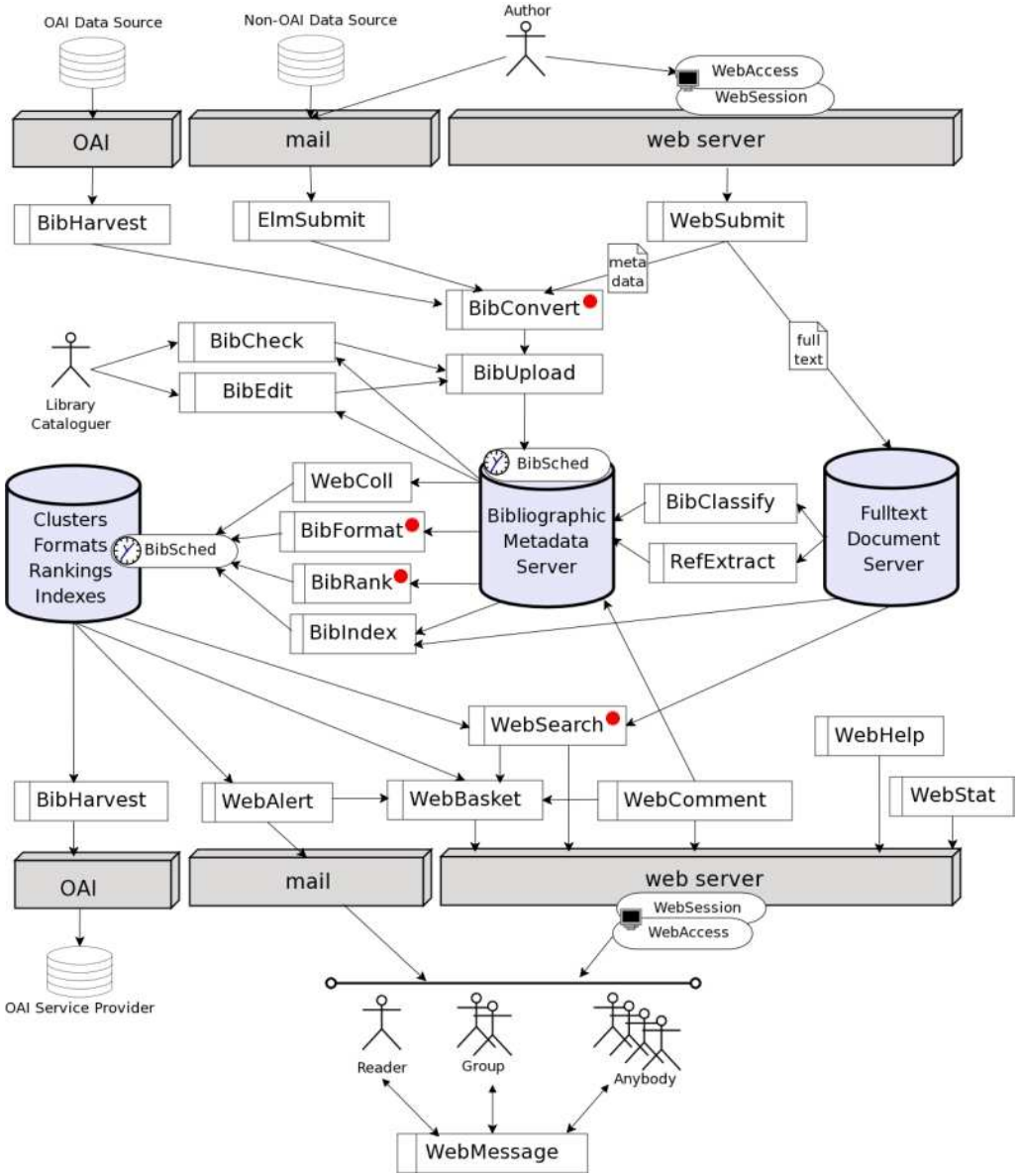


CDS Invenio demo site in Greek, one of 20 available UI languages

Inspire Phase IIa (May 2008)

- *goal*: reproducing existing *user-level* functionalities
- SPIRES **data conversion** and load
  - ⋆ mapping to MARC
  - ⋆ load 760,000 literature and institute records
- SPIRES **citation analysis**
  - ⋆ self-cite elimination
  - ⋆ cite summary and related formats
- SPIRES **search syntax**
  - ⋆ supporting SPIRES traditional syntax
  - ⋆ fuzzy author search
- **output formats** and **site layout**
- . . . May 2008: public release for alpha testers?

Phase IIa: Invenio/Inspire modules



Invenio software:

- 24 modules
- 160,000+ Python LOCs
- 470+ test cases

Inspire Phase IIa extensions:

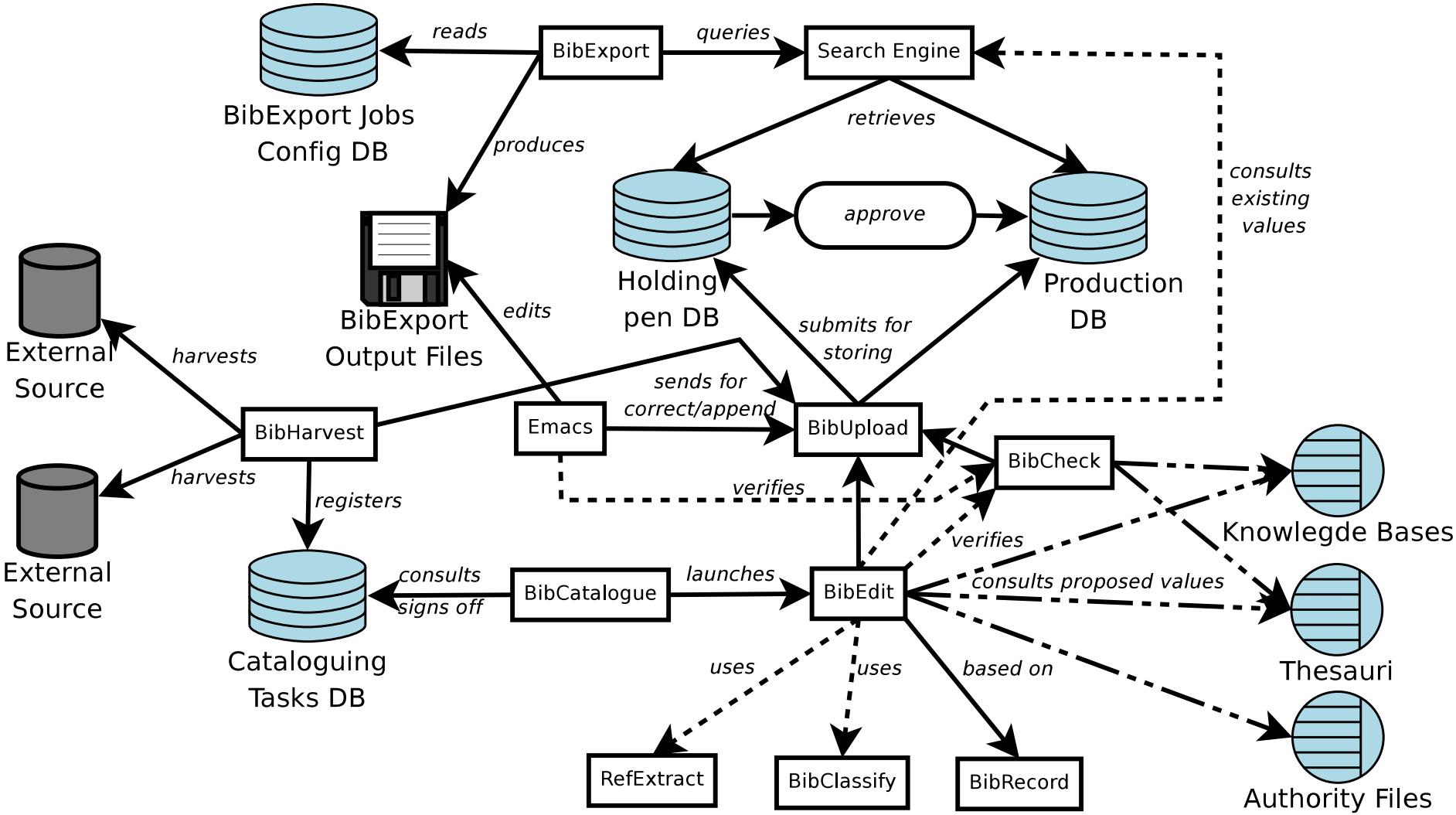
- BibConvert extensions (SPIRES to MARC mapping)
- BibRank extensions (cite summary)
- WebSearch extensions (SPIRES search syntax)
- BibFormat, WebStyle (output style)

Inspire Phase IIb (Q2 2009)

- *goal*: reproducing existing *cataloguer-level* functionality
- record editing interface
- record checking tools
- record maintenance tools
- record inputting workflow
- record harvesting workflow
- knowledge bases

- . . . Invenio traditionally relied on ALEPH cataloguing tool
- . . . building strong **native cataloguing tools** for Inspire
- . . . Q2 2009: public release of the production service?

Phase IIb: Principal Developments



Inspire Phase III (from 2009)

- *goal*: *adding value* on top of reproduced SPIRES functionality
- *__user accounts__ (even though most users are guests)
  - *local accounts? authentication specific to every lab? *single sign-on and *certificates?
  - e.g. CDS: 7,709 registered users, 67% non-CERN
- *__collaborative features__
  - *user groups
  - *sharable baskets
  - *notification alerts
  - *user recommendations
- *__keyword taxonomy__, user __tagging__

(* = already present in CDS Invenio)

Inspire Phase III (from 2009) cont'd

- outspiring: *reaching outside* Inspire community & HEP
  - ⋆ other related fields and open source communities
  - ⋆ relations with commercial publishers
- **big brother author database**
- big brother institute/conference/experiment databases
- *__full-text file__ treatment, OA, publisher negotiations
- advanced **linguistic context** indexing
- advanced content indexing of **plots** and **tables**
- *__conference contributions__, crawling static web sites
- extended **citation networks**, combined **impact metric**
- open **data mining** and **bibliometric** studies

(∗ = already present in CDS Invenio)

Conclusions

- the SPIRES collaboration and the CDS Invenio open source community join forces
- unique opportunity to build *the* single stop shop HEP information system
- inspired by HEP needs, inspiring other disciplines
- a win-win situation for our institutes
  - ⋆ CERN, DESY, Fermilab, SLAC
  - ⋆ . . . interesting for other partners?
- more information:

  <http://hep-inspire.net/>
  <https://twiki.cern.ch/twiki/bin/view/Inspire>