INVENIO @ CERN

Jérôme Caffaro

Software engineer in IT-UDS-CDS

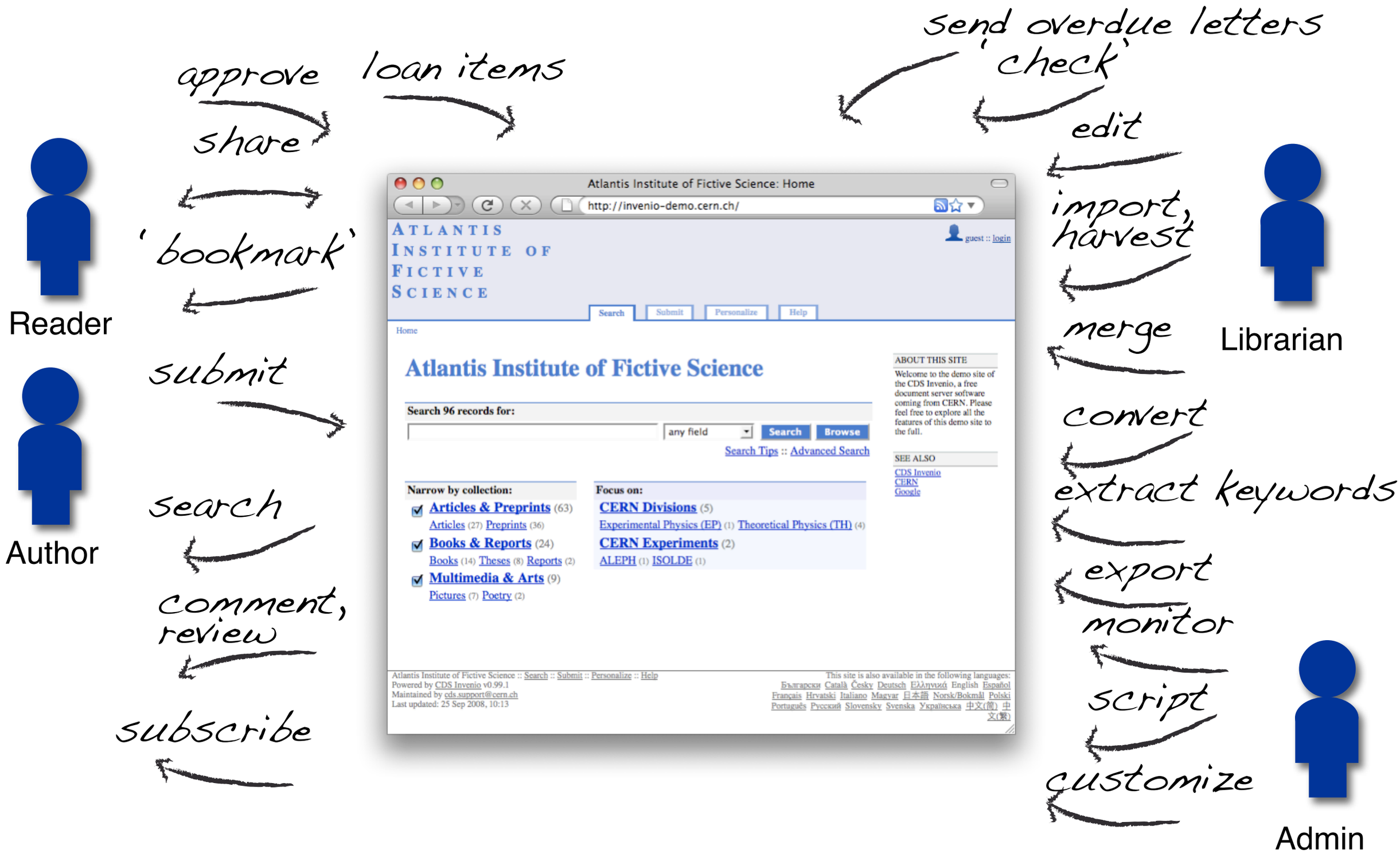
- What is Invenio?
- The CERN Document Server (CDS)
- Integration with other services
- Development strategy
- Current challenges & glimpse at the future

> Update on INSPIRE

What is Invenio?

- Integrated Digital Library / Repository software
- A platform of choice for managing documents in HEP
  - also adopted in other fields (medium to big repositories)
- Web application
- Open-source GPL project
- Python (mostly), MySQL and Apache
- Based on open standards
  MARCXML, OAI-PMH, OpenURL, OpenSearch, etc.
- Flexible, scriptable

What is Invenio?



Early History

1954 CERN library starts paper dissemination of preprints

1965 First computers at CERN library to help with cataloging

1990 Electronic distribution of preprints via FTP

1993 CERN Preprint Server, web front-end of electronic preprint catalogue. Institutional repository

1996 CERN Library Server (weblib): added books, periodicals and "other material".

2000 CERN Document Server: multimedia material, internal notes)

2002 First public release of the software under GNU-GPL. Worldwide installations and collaborations

Early History

1954 CERN library starts paper dissemination of preprints

1965 Firs N library to help with cataloging

1990 Electronic

1993 CERN Pr c preprint catalogue. Institutional repository

1996 CERN Library Server (weblib): added books,

2000 CERN Document Server: multimedia material, internal notes

2002 First publi e u Worldwid ora

1989: the web is born

1991
- SPIRES (SLAC) First database available on the web
- arXiv.org as repository for particle physics

2007: Started collaboration with SPIRES

~2008: Contacts with NASA ADS

~2010: Contacts with arXiv.org

↦**2011:** Adopted by ~30 institutions worlwide

Invenio installations at CERN

BlogForever
*just started

CDS | INSPIRE | ILCDoc | CDS adm | OpenAire
"orphan" repo

INVENIO @ CERN

CERN Document Server

CERN Document Server

CDS | Indico | Bibliothèque | Bulletin | EDMS

Recherche | Soumission | Aide | Your CDS | identification

Accueil > Books & Proceedings > Books

Books

Chercher dans 46,464 notices:

tous les champs | Recherche | Liste

Conseils de recherche :: Recherche avancée

Derniers ajouts:

2010-11-20 06:00 — **The Oxford style manual** / Ritter, R.M (ed.)
Oxford : Oxford Univ. Press, 2003. - 1033 p.
This book at Amazon
Notice détaillée - Notices similaires

2010-11-19 09:13 — **Microsoft office : SharePoint server 2007 : implémentation en environnement SQL server 2008 et Windows server 2008** / Bories, William
Paris : Éditions ENI, 2009. - 358 p. CERN library copies
This book at Amazon
Notice détaillée - Notices similaires

2010-11-19 09:13 — **SharePoint 2007 : créez votre site collaboratif étape par étape** / Schmitt, Sandrine
Paris : Éditions ENI, 2008. - 274 p. CERN library copies
This book at Amazon
Notice détaillée - Notices similaires

2010-11-19 09:13 — **Ouverture** / Cahin, Gerard (ed.) ; Poirat, Florence (ed.) ; Szurek, Sandra (ed.)
Paris : Editions A. Pedone, 2007. - 389 p. (La France et le droit international ; I)

Focaliser sur:
CERN Bookshop (1,382)
English Book Club (54)

Rechercher également dans:
KISS Books/Journals

CERN Document Server

CERN Document Server

CDS | Indico | Bibliothèque | Bulletin | EDMS

Recherche | Soumission | Aide | Your CDS | identification

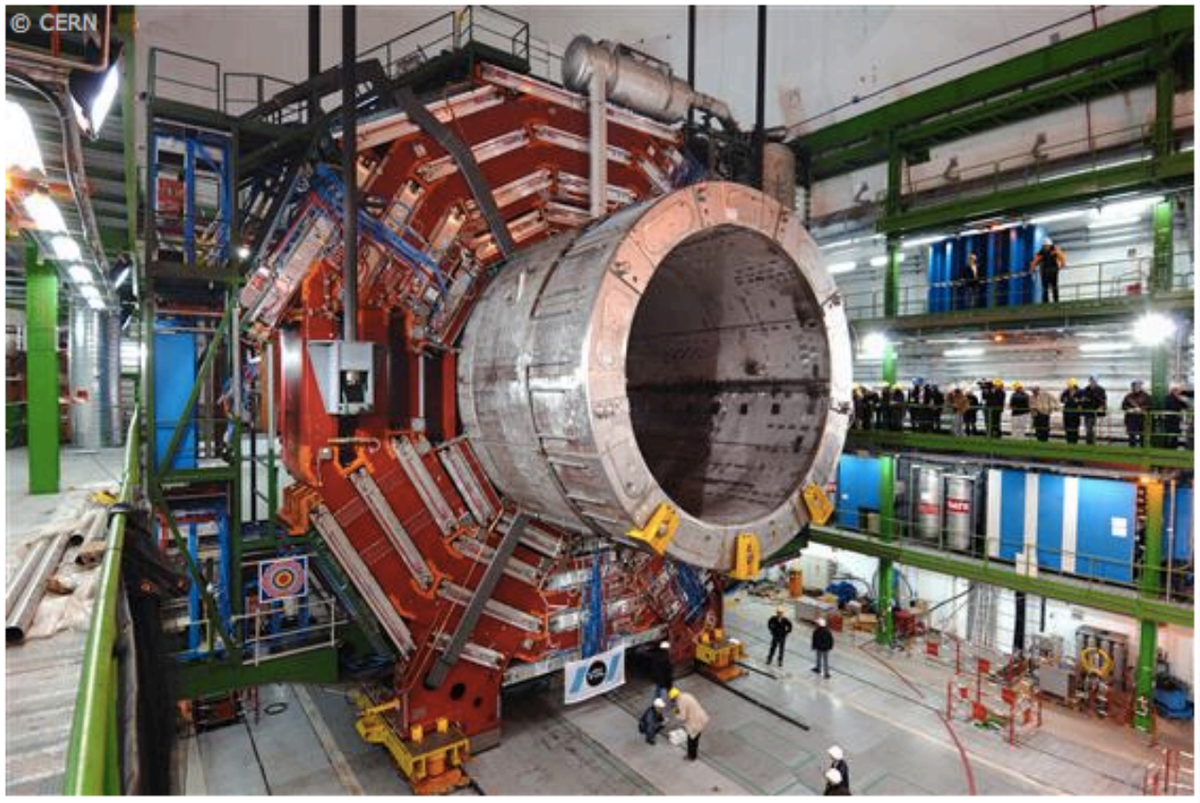
Accueil > Multimedia & Outreach > Photos > CERN PhotoLab > Notice#1020311: The central part of CMS is lowered

Informations | Références | Discussion | Fichiers

CERN PhotoLab — CERN-EX-0702022

The central part of CMS is lowered

Descente dans la caverne de la partie centrale du détecteur de particules CMS (Compact Muon Solenoid).



haute résolution?

2022 03
High-res

CERN-EX-0702022 04
Small Medium Large High-res

CERN Document Server

CERN Document Server

CDS | Indico | Library | Bulletin | EDMS

Search | Submit | Help | Your CDS | login

Home > Articles & Preprints > Preprints > Information management

Information | Discussion | Files

Preprint

Report number	CERN-DD-89-001-OC
Title	**Information management : a proposal**
Author(s)	Berners-Lee, Timothy J (CERN)
Imprint	Mar 1989. - 15 p.
Subject category	Particle Physics - Theory

Record created 1998-10-28, last modified 2010-11-11

Similar records

Access to fulltext document:
PDF
PS.GZ

Add to personal basket
Export as BibTeX, MARC, MARCXML, DC, EndNote, NLM, RefWorks

CERN Document Server

CERN Document Server

CDS | Indico | Bibliothèque | Bulletin | EDMS

Recherche | Soumission | Aide | Your CDS | identification

Accueil > Multimedia & Outreach > Videos > Video Movies > Notice#1270161: Library Induction Clip

Informations | Références | Discussion | Fichiers

Library Induction Clip
Presentation of Library for new arrivals

© CERN

00:00 00:00

Windows Media Flash

Download Movie

Flash:		High (753 kbps)
Windows Media:	Medium (480 kbps)	High (753 kbps)

Download high-res version: mpg (201.0 MB), mov (830.0 MB)
Need help to download high-resolutions?

A short clip introducing the library - used in the induction process for those new to CERN.

Produced by: CERN Video Productions
5 min 20 s min. / 05 June 2010 / **CERN AV**
Keywords: CERN Library, Library, Induction, Induction video, Presentation of library

Language: English
Source Medium: DVCproHD PAL

Reference: CERN-MOVIE-2010-117

Voir également:

- Clip CERN Director General Rolf Heuer looks forward to first physics at the LHC
- LHC first physics : clip resume of the day March the 30th, 2010
- Recruitment Clip Come to work at CERN
- Computer Security at CERN
- Clip (version française) Le directeur général du CERN Rolf Heuer attend avec impatience les collisions à 7 TeV avec le LHC
- Computer security at CERN English subtitled version
- Le CERN programme un premier essai de

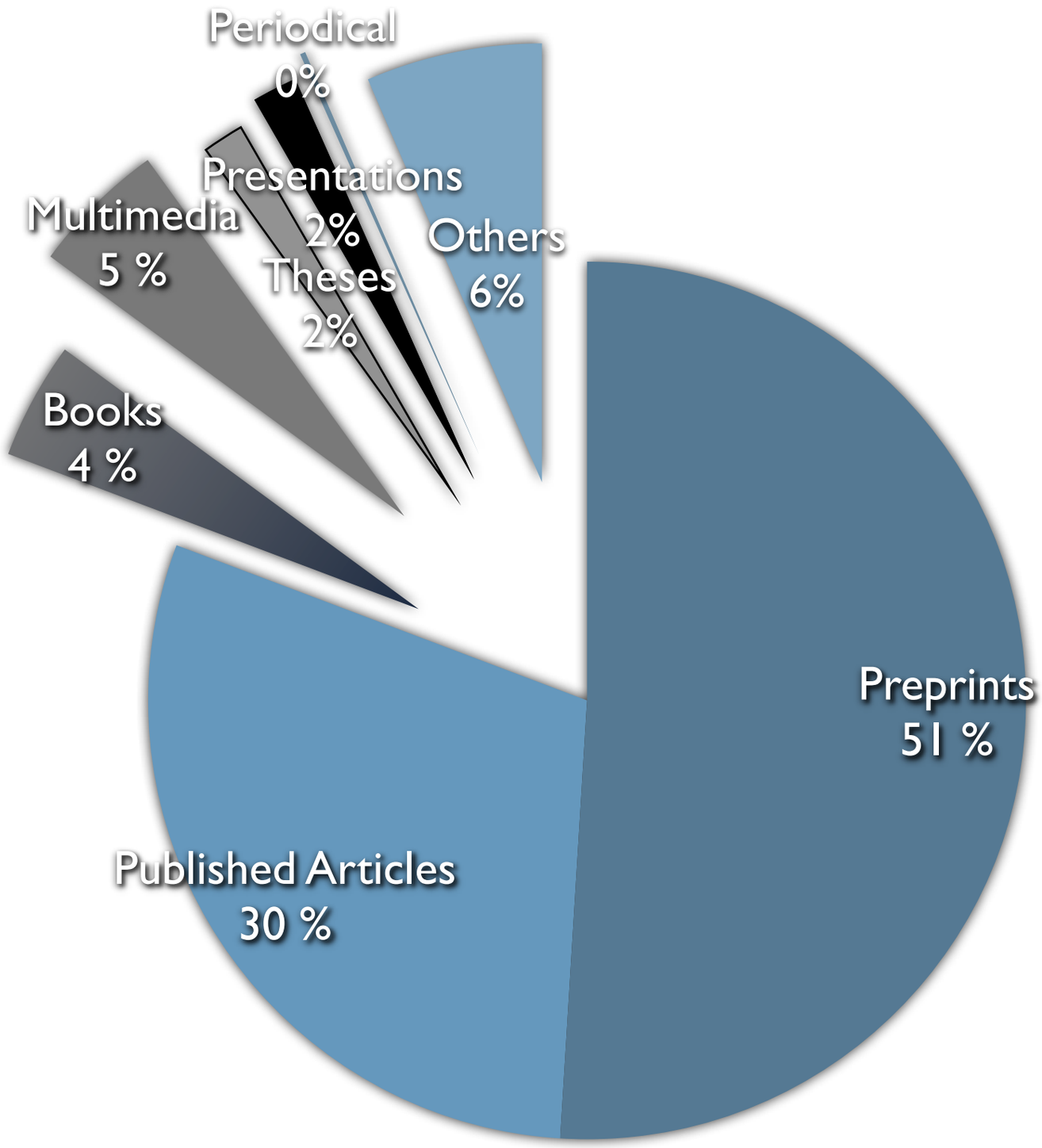
CERN Document Server

CERN Document Server

- > 1M records, > 700 collections
- > 150 bibliographic formatting templates
- > 100 submission workflows (+ sub-workflows)
- Variety of document types: books, video lectures, preprints, photos, reports, etc.
- Institutional and subject-based repository
- ~18k search queries per day*
- >200 new documents per day*

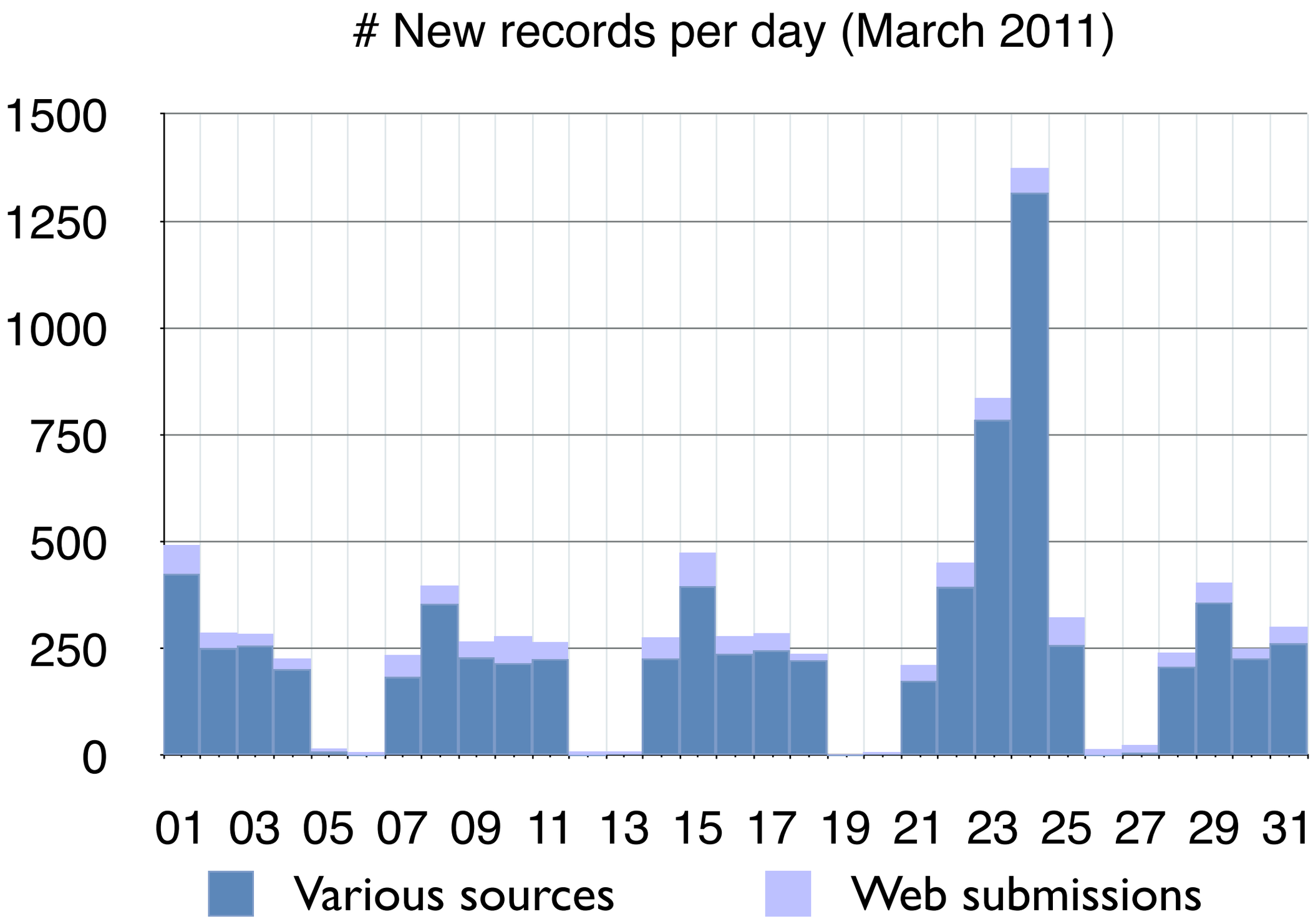
Data: April 2011 * Average including week-ends

CERN Document Server

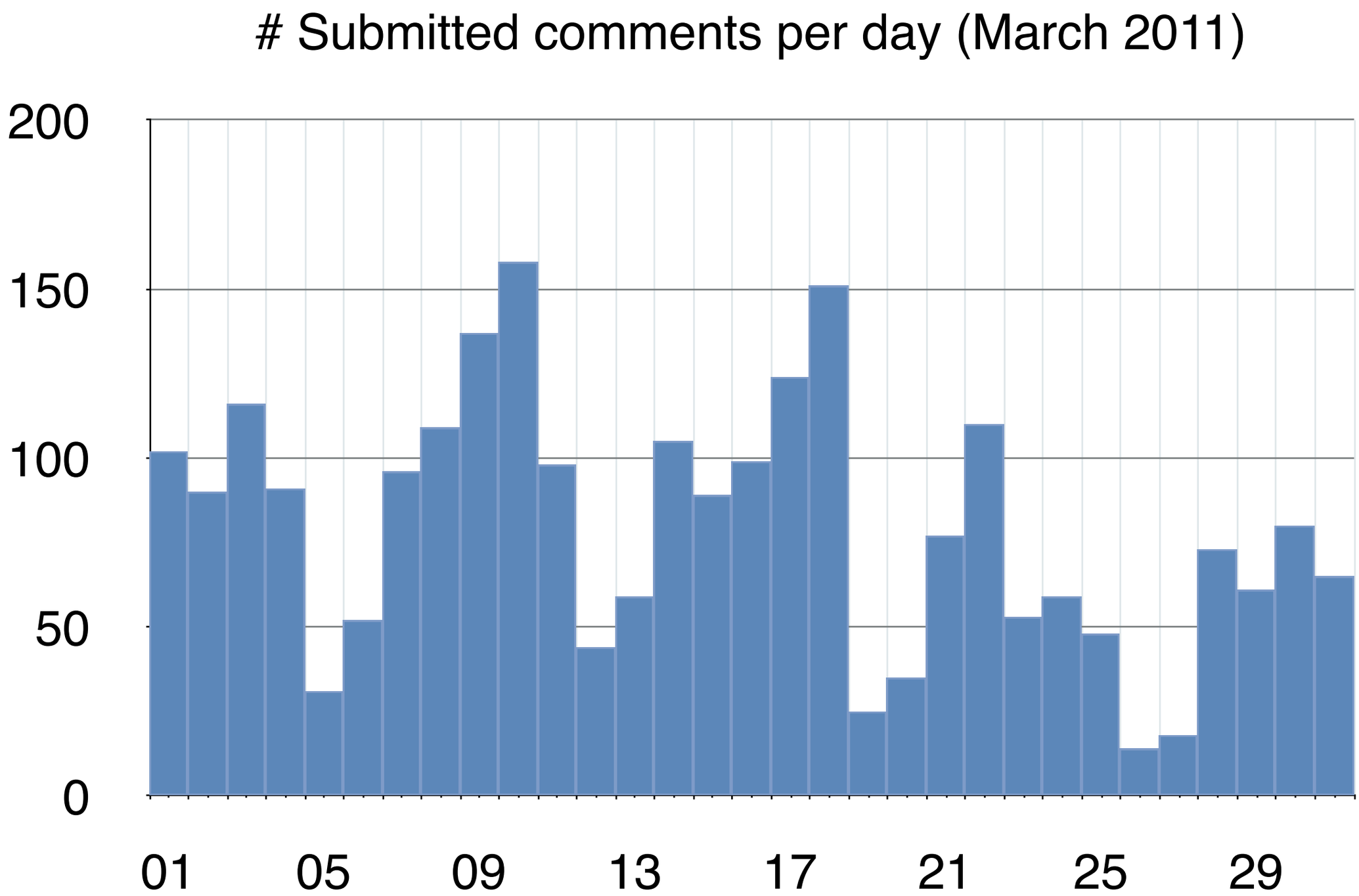


Data: 2nd November 2010

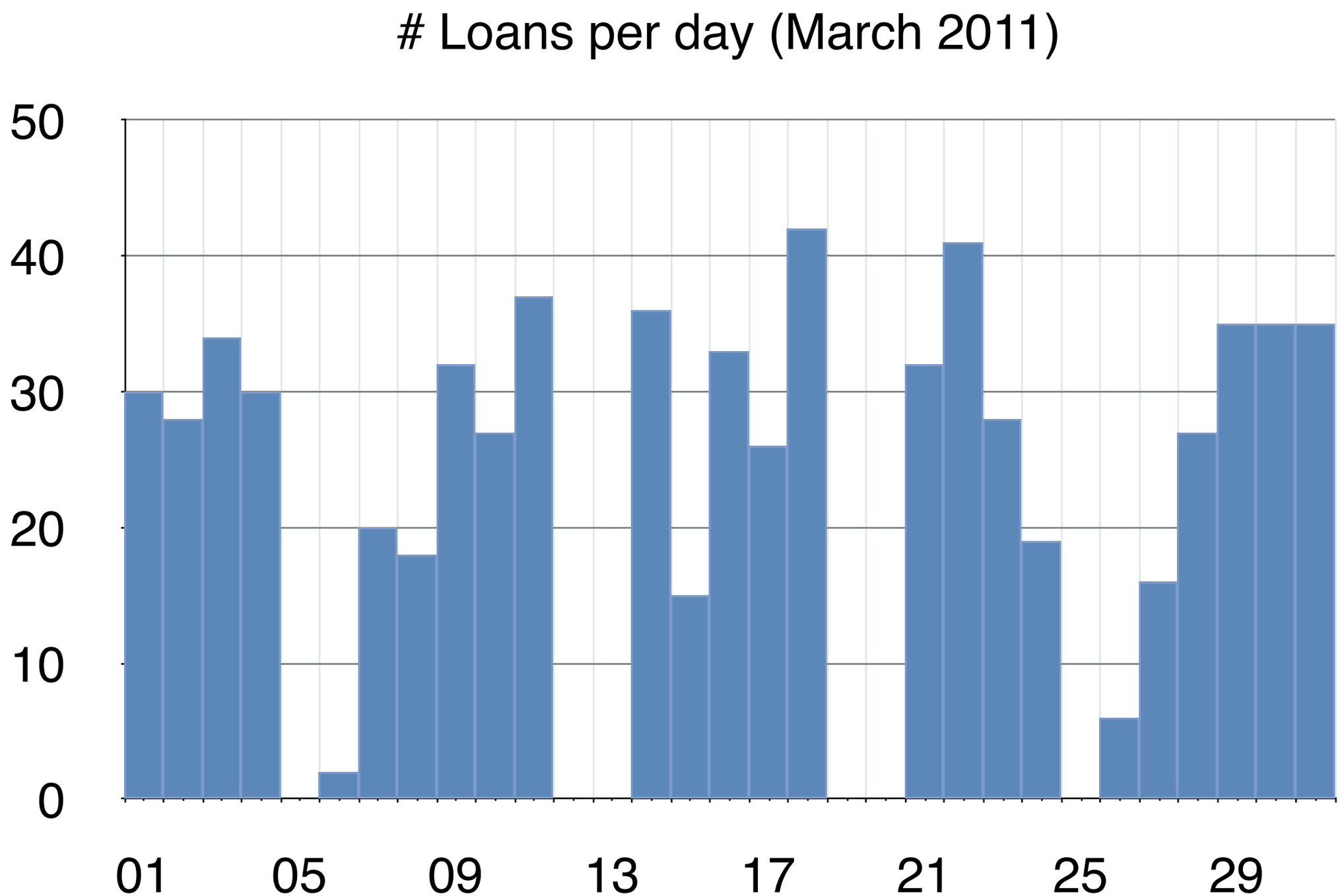
CERN Document Server



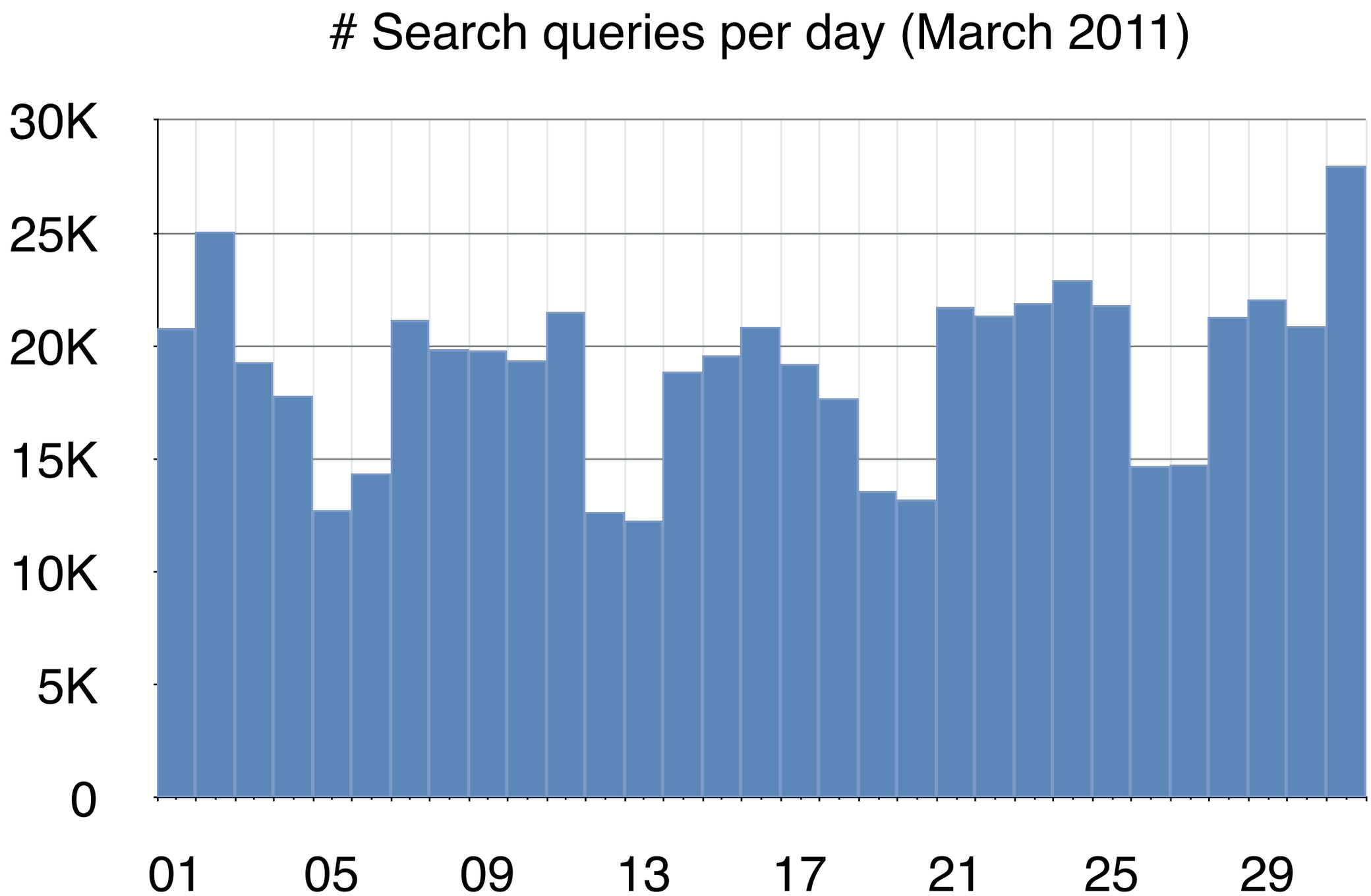
CERN Document Server



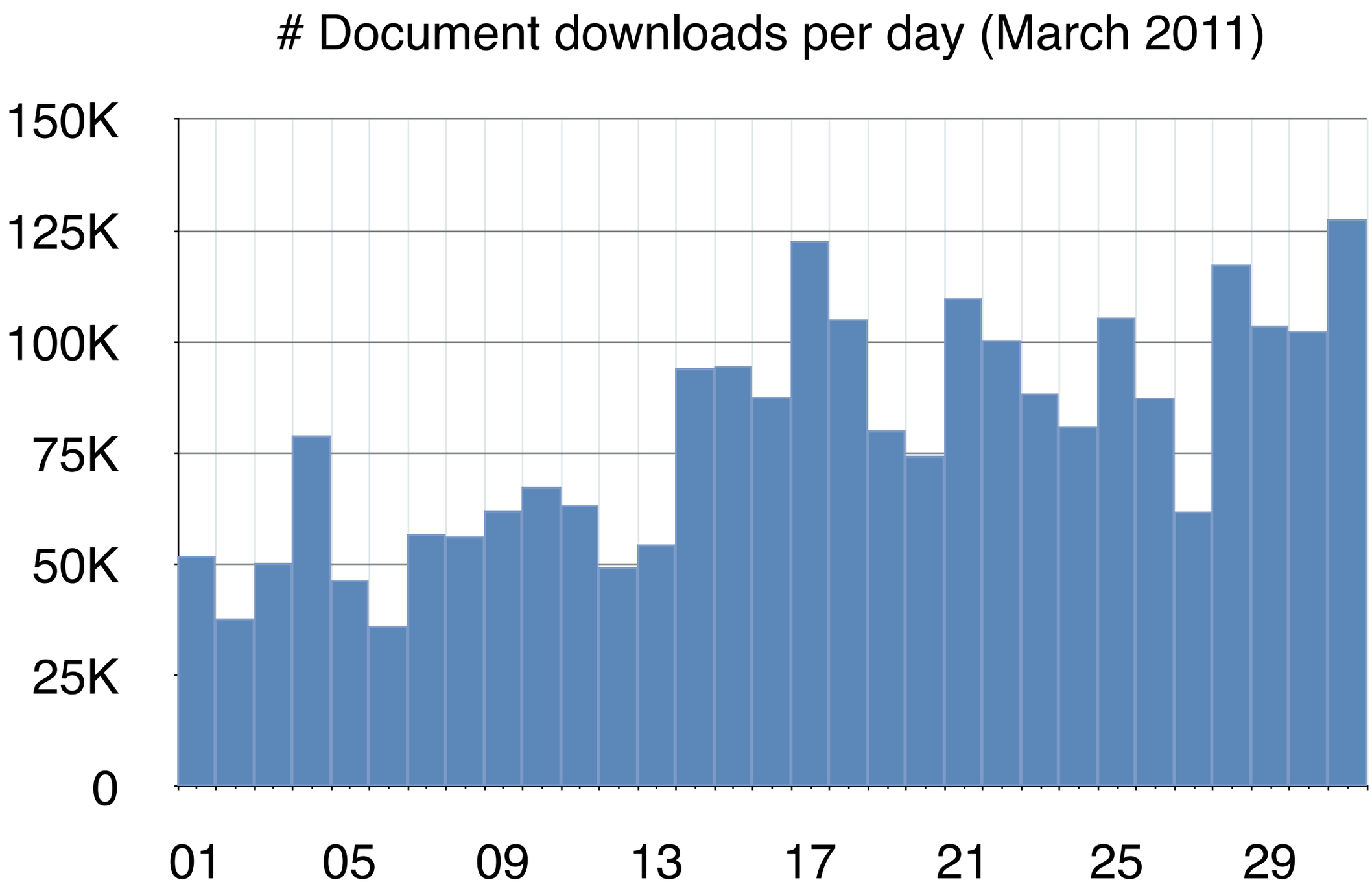
CERN Document Server



CERN Document Server



CERN Document Server



CERN Document Server

- **2,417 alerts**
  - ... set up by **1,615** users (22% from CERN)
  - ... some alerts going to large mailing lists
- **6,245 baskets**
  - ... set up by **4,575** users (27% from CERN)
  - ... **481** being shared (various access rights)
  - ... covering **123,211** records

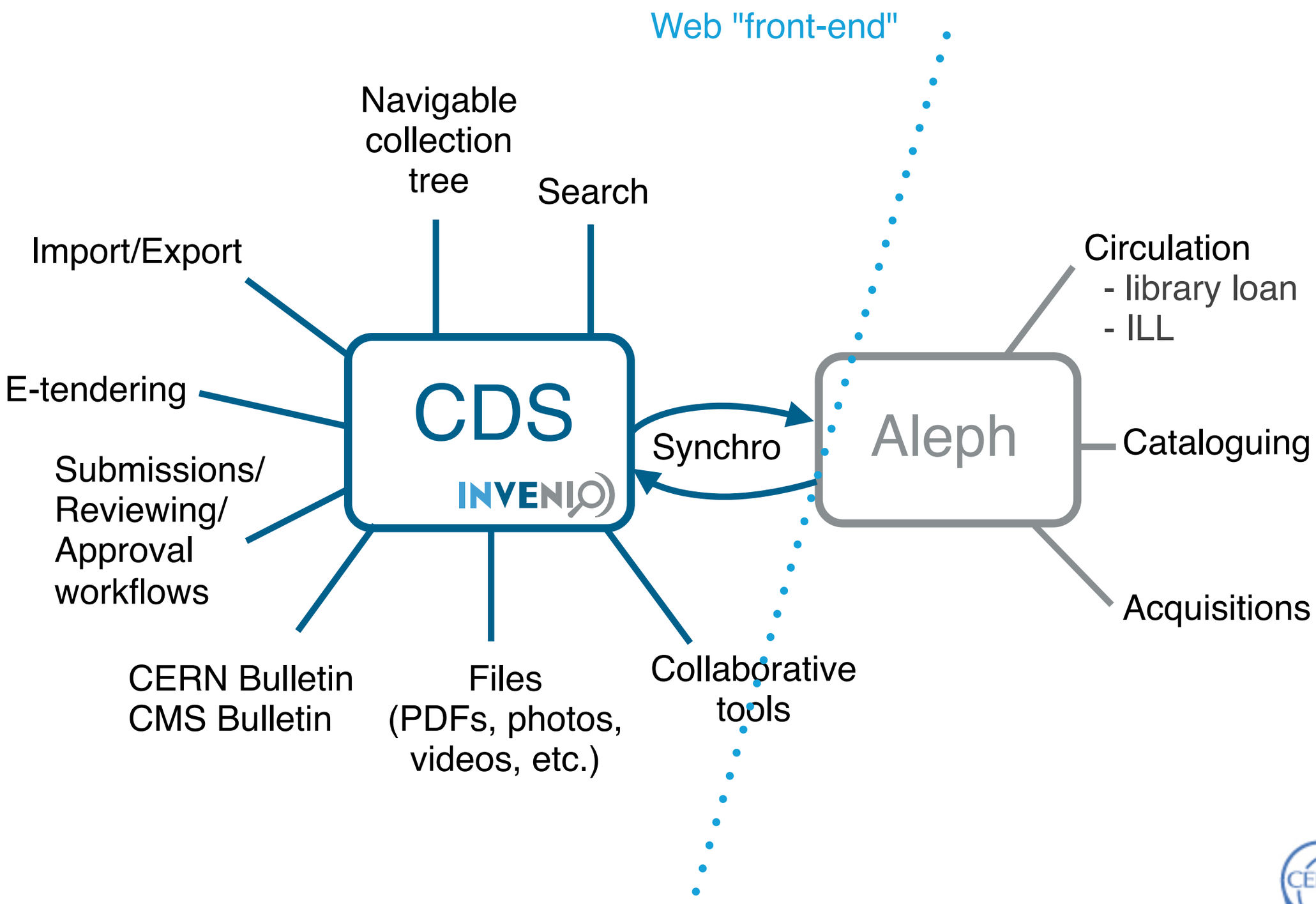
Data: November 2010

Invenio at CERN

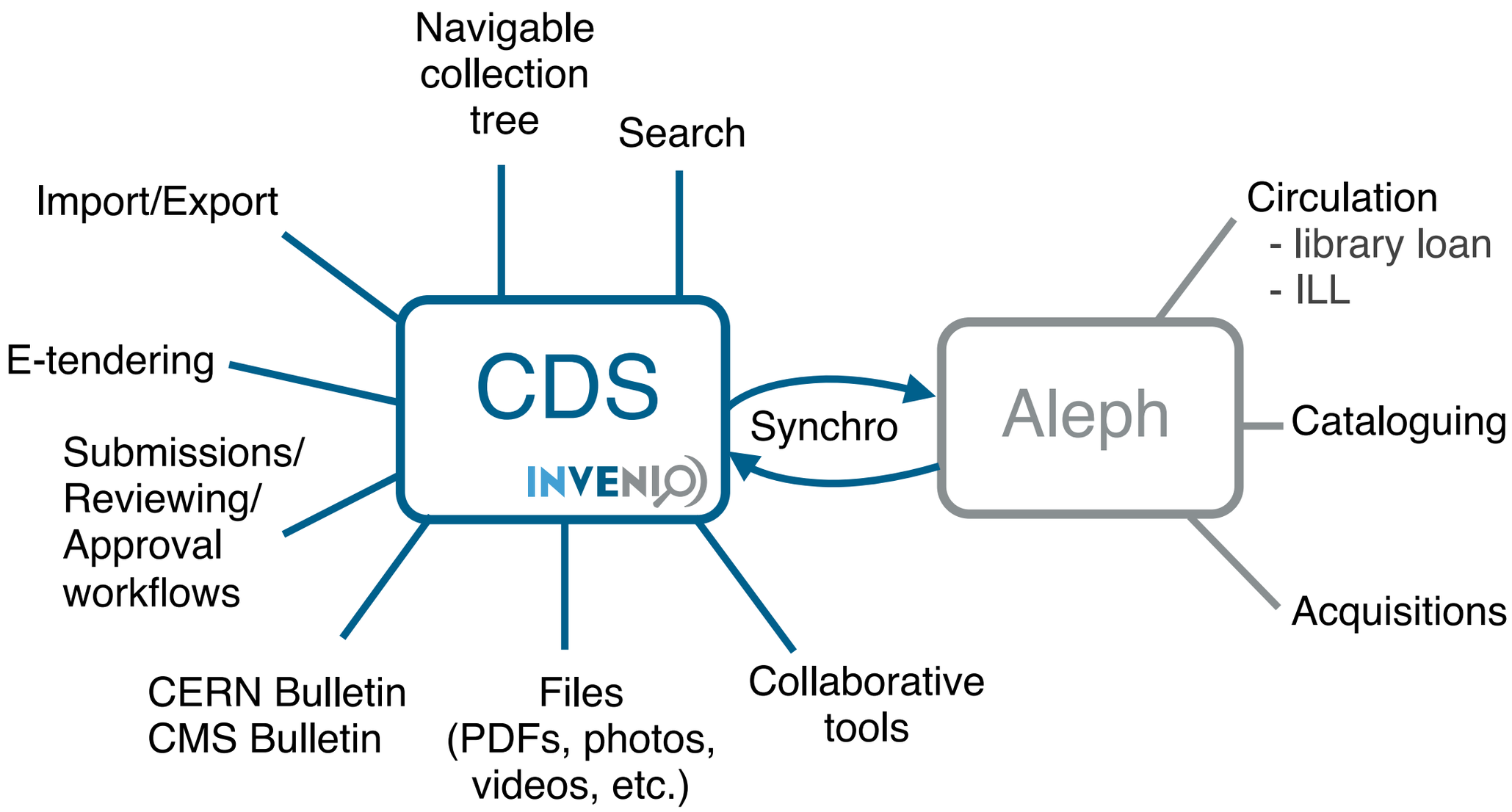
CDS

INVENIO

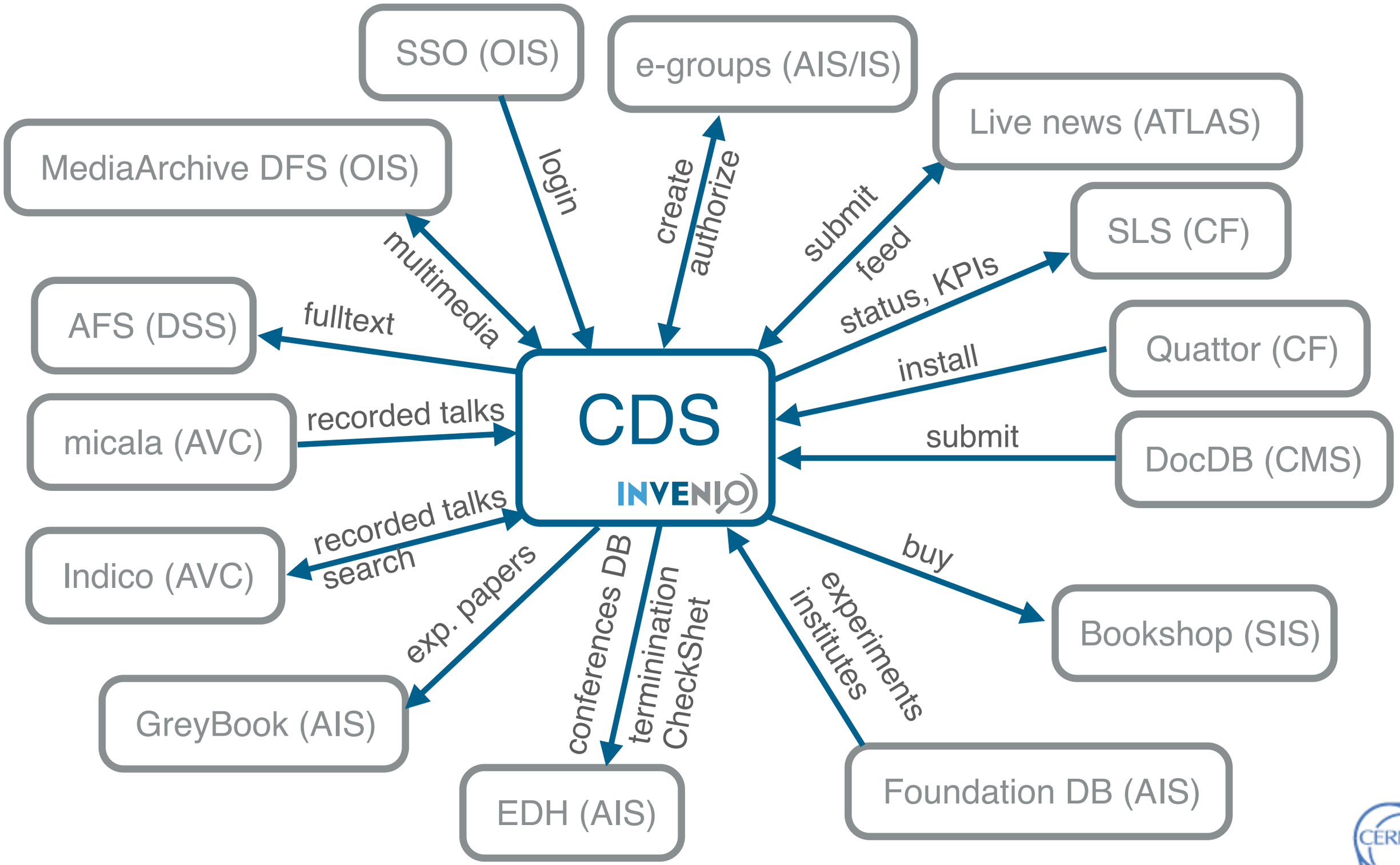
Invenio at CERN



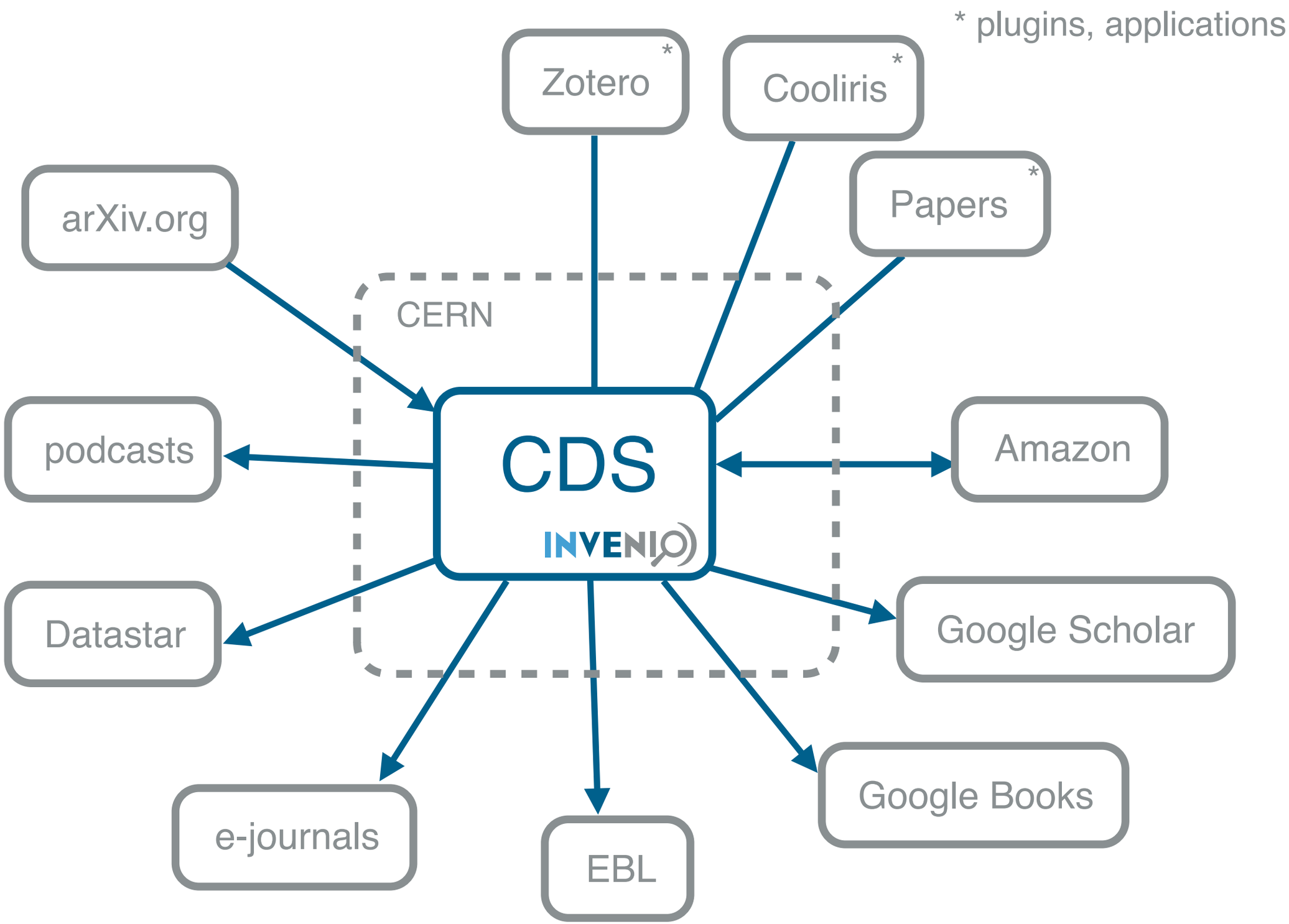
Invenio at CERN



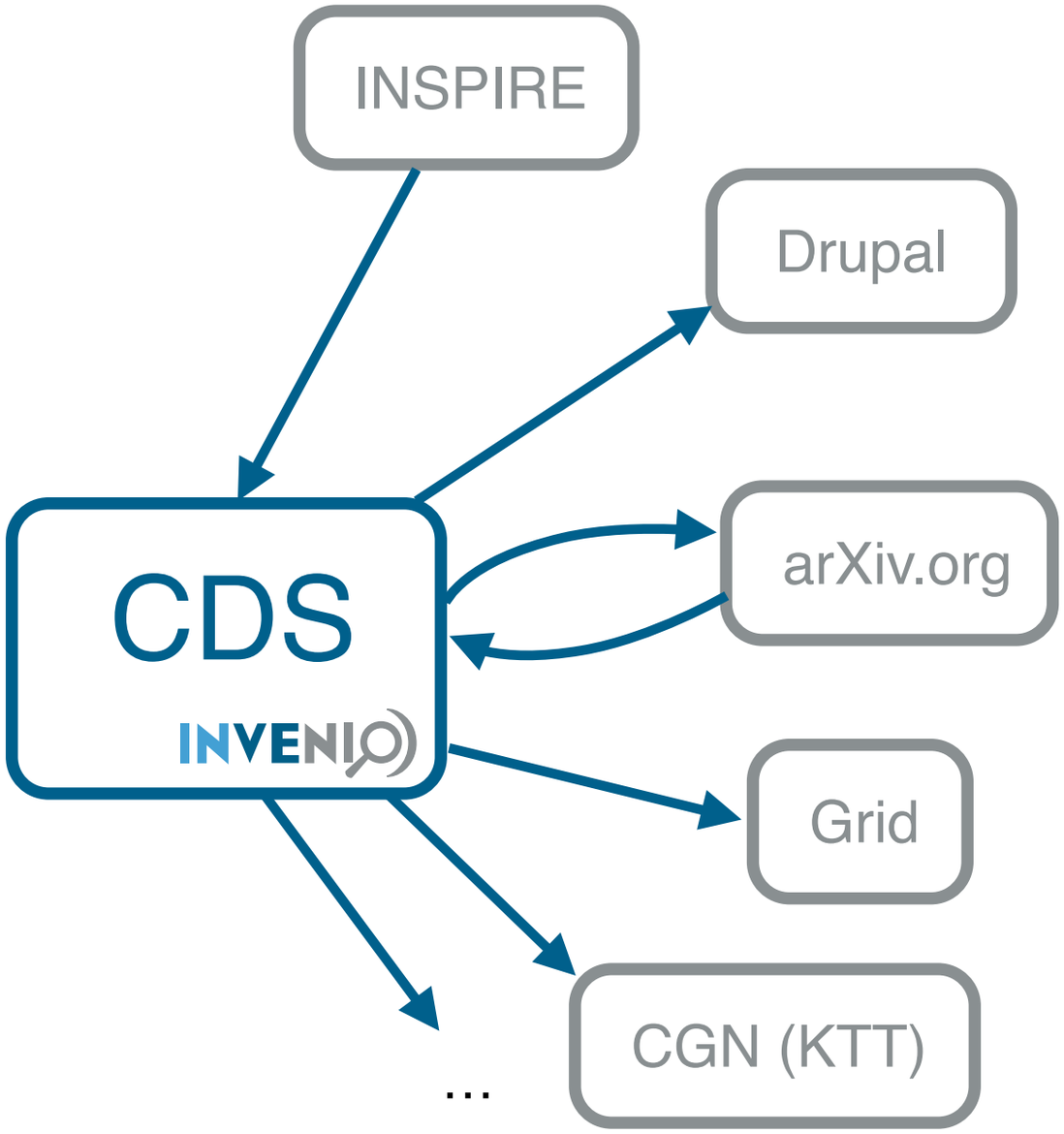
Invenio at CERN



Invenio at CERN

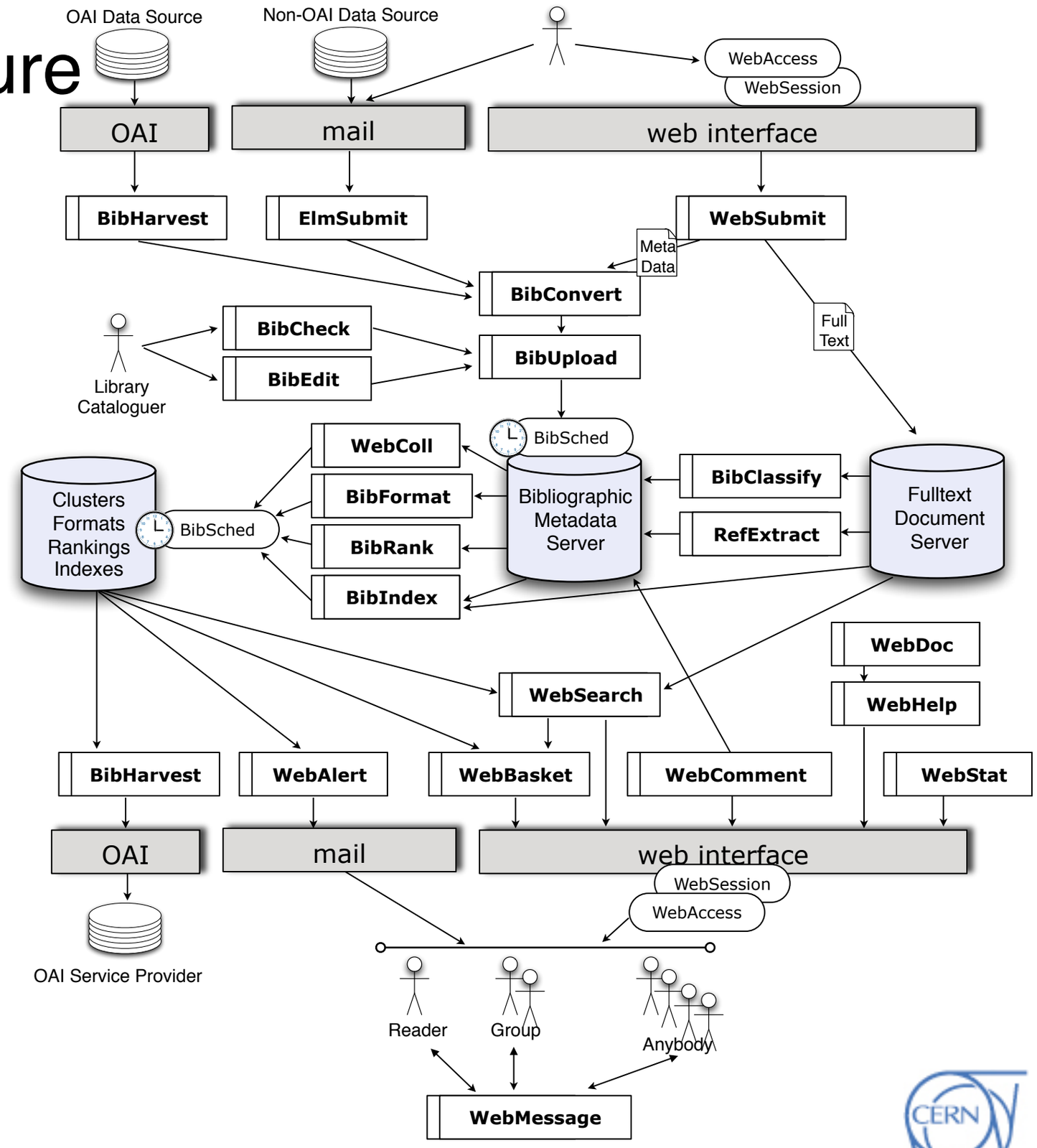


New/Future integrations



Interoperability made possible

- Modular architecture



Interoperability made possible

- Modular architecture
  - building blocks for customized workflows
  - easily "overridable/upgradable" components
- Use of standards data and communication protocols
- CLIs & Pythonic/Web APIs
  - Rapid prototyping with Python
  - Simple REST interfaces
- Plugin-enabled components
- Submission workflow engine

Hardware & Software

- **CDS**
  - Quattor-managed SLC5 machines, some in-house prepared RPMs
  - **2x** Dell Blade, Intel(R) Xeon(R) CPU E5410 @ 2.33GHz (4 cores), 16 GB RAM
- **INSPIRE**
  - Quattor-managed SLC5 machines
  - **4x** Dell Blade, Intel(R) Xeon(R) CPU L5410 @ 2.33GHz (4 cores), 16 GB RAM
    - 1x DB
    - 2x workers
    - 1x load balancer

Development strategy

- **~ 33 modules**
  - ~ 290,000 lines of Python code
  - ~ 30,000 lines of HTML
  - ~ 12,000 lines of Javascript
  - ~ 6,000 lines of XSL
- **organic-growth** software development model
- ~ 80 contributors (over ~10 years)
  - many **temporary members (students, associates, etc.)**
- version control system: **Git**
  - good for distributed teams
  - interplay with **SVN**

Quality Assurance

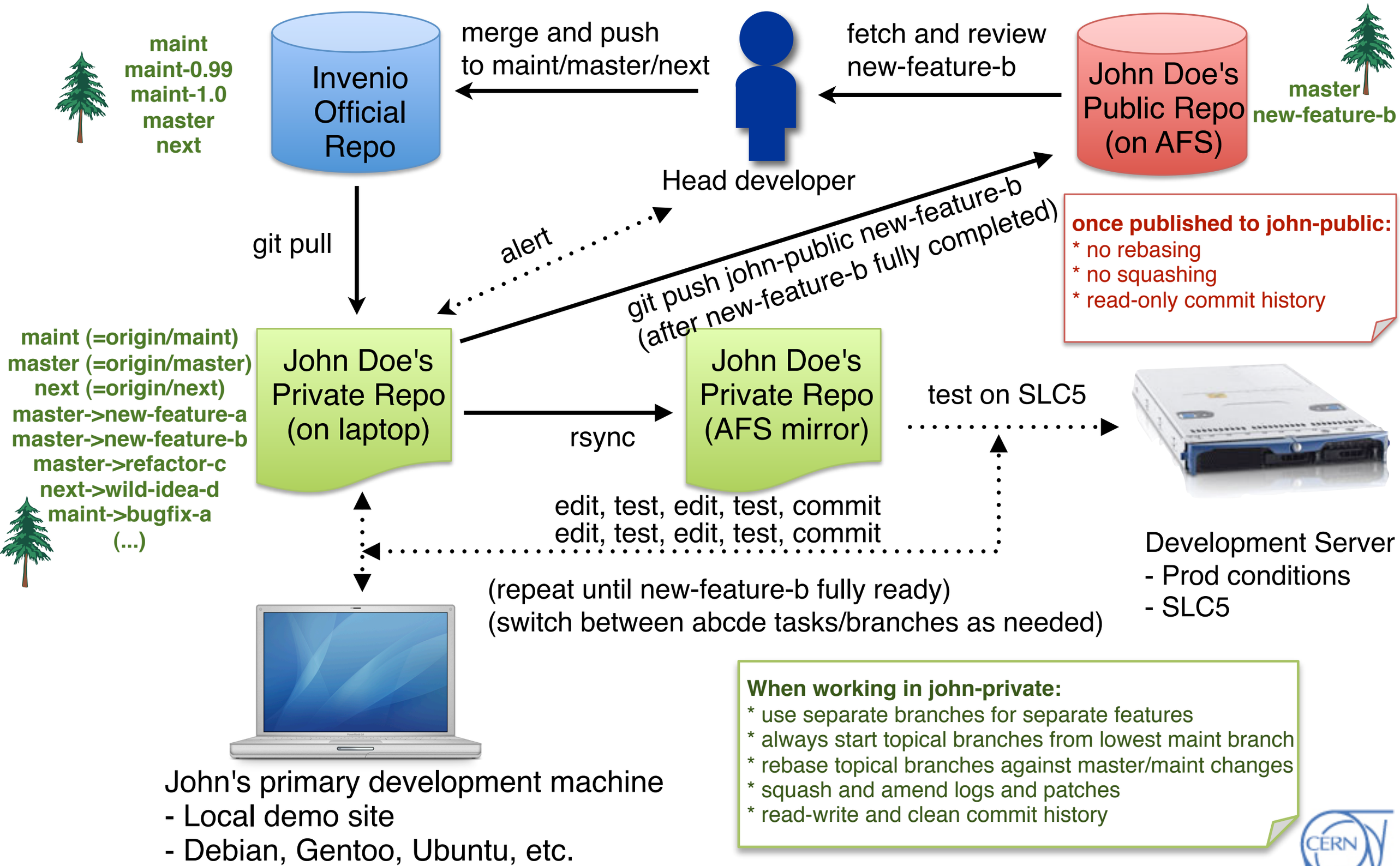
- **Coding standards**
  - Eg. PEP8 (Style Guide for Python), etc.
- **Documentation**
  - *"If the code and the comments disagree, then both are probably wrong."* – attributed to Norm Schryer
- **Test** suite
  - ~1,000 unit/regression/web tests
- **Security**
  - XSS, CSRF, SQL injection, etc.
- Code **review**
- **Kwalitee** check: "measuring" quality
  - *"It looks like quality, it sounds like quality, but it's not quite quality."* – CPAN Testing Service (quoting Michael Schwern)

Kwalitee

Module	#LOC	#UnitT	#RegrT	#WebT	#T/1kLOC	#MissDoc	#PyChk/1kSRC	PyLintScore	PyLintDetails
bibauthorid	15664	5	0	0	0.32	27	5.107	9.18/10	139F 2E 149W 161R
bibcatalog	641	0	0	0	0.00	1	9.360	6.47/10	0F 0E 11W 18R
bibcheck	338	0	0	0	0.00	0	2.959	7.31/10	0F 0E 6W 1R
bibcirculatio	29102	0	3	0	0.10	30	1.787	4.36/10	1F 2E 16W 314R
bibclassify	3692	0	2	0	0.54	3	2.709	9.40/10	21F 0E 46W 29R
bibconvert	2963	18	3	0	7.09	2	37.462	7.48/10	2F 2E 91W 21R
bibedit	13979	67	2	4	5.22	4	2.647	7.64/10	0F 14E 85W 123R
bibexport	4252	1	0	0	0.24	0	3.763	8.51/10	0F 0E 38W 31R
bibformat	16083	30	20	1	3.17	2	6.156	8.31/10	0F 0E 275W 133R
bibharvest	7185	3	11	1	2.09	66	17.676	6.30/10	0F 2E 141W 120R
bibindex	5800	43	2	0	7.76	36	16.034	6.74/10	2F 2E 159W 65R
bibknowledge	2753	0	10	0	3.63	0	2.906	6.39/10	1F 10E 14W 38R
bibmatch	1001	0	9	0	8.99	0	5.994	8.15/10	0F 0E 9W 10R
bibmerge	1649	0	0	0	0.00	12	7.884	7.57/10	1F 0E 84W 27R
bibrank	7814	9	14	0	2.94	44	14.717	6.93/10	0F 1E 253W 91R
bibsched	2534	0	0	0	0.00	31	5.919	7.97/10	0F 0E 47W 26R
bibsword	5914	0	0	0	0.00	1	3.889	8.80/10	0F 9E 18W 60R
bibupload	3689	0	58	0	15.72	6	18.975	7.74/10	0F 0E 107W 62R
elmsubmit	6013	3	0	0	0.50	135	14.801	6.03/10	15F 31E 168W 49R
miscutil	13211	192	9	0	15.21	38	6.510	8.09/10	8F 9E 192W 104R
webaccess	8522	26	10	0	4.22	12	18.540	6.85/10	2F 0E 156W 100R
webalert	2456	3	2	3	3.26	12	3.664	7.35/10	0F 0E 32W 30R
webbasket	10776	0	1	0	0.09	8	4.640	2.14/10	0F 1E 136W 177R
webcomment	6033	2	8	3	2.15	3	11.437	5.12/10	1F 29E 173W 98R
webjournal	7930	2	38	1	5.17	1	5.422	7.70/10	0F 0E 149W 103R
webmessage	2570	4	19	4	10.51	1	1.556	8.58/10	0F 0E 32W 35R
websearch	20663	163	158	1	15.58	91	14.519	7.13/10	1F 3E 585W 346R
websession	8172	3	10	4	2.08	50	11.992	7.27/10	0F 0E 130W 118R
webstat	6406	0	1	0	0.16	6	3.590	7.55/10	4F 0E 33W 44R
webstyle	5387	9	2	0	2.04	68	16.336	6.88/10	0F 3E 150W 78R
websubmit	46555	0	17	6	0.49	295	15.229	2.70/10	11F 197E 1364W 497
TOTAL	269747	583	409	28	3.78	985	9.668	7.05/10	209F 317E 4849W 310

*development snapshot 1.0.0-rc0.283-70380

Git collaboration model



Examples of what's next

- **Software-wise**
  - **Examples of forthcoming features**
    - Wider Solr integration
    - New generation submission workflow
    - Improved search interface (Facets, add-to-search, restricted collections, etc.)
  - **Stability and consolidation**
    - Expand load-balancing (mod_proxy_balancer)
- **Service-wise**
  - Drupal
  - CDS ↔ INSPIRE ↔ ADS ↔ arXiv
  - Automated deployment from Git

Race for physics results

- Data is flowing... Results are produced
- Impact on Invenio and CDS
  - Confidentiality / access restrictions
  - Authorship
  - Authoring tools
  - Copyright
  - Immediacy
  - Various, evolving policies

Finally...

- CDS provides key service for the publication of physics results in close collaboration with the experiments publication committees.
- CDS is CERN institutional repository:
  - Archive: preserving history
  - Outreach: access to most recent publications, images, etc.
- Growing adoption of Invenio
  - At CERN, in HEP, EU projects and other disciplines.
  - Larger community of developers. More collaborations.

Some news on INSPIRE

- "Claim your paper" feature
  - Crowdsourced
  - Author disambiguation
- Minor UI changes
- Improved SPIRES search syntax compatibility
- Automated arXiv harvesting
- Solr integration (fulltext search)
- Plot display and caption search

Q & A

http://invenio-software.org

http://cds.cern.ch

http://inspirebeta.net