Realizing the Dream of a Global Digital Library in High-Energy Physics

Annette Holtkamp, Salvatore Mele, Tibor Simko, Tim Smith
CERN, Geneva

DML 2010 – Paris 7 Jul 2010
HEP community

- closely-knit community
  - 20-30k active researchers publishing 10k articles
  - large collaborations (up to 5000 members)
  - very international (even small author groups)
  - authors = readers

- rapid information exchange essential
  - mailing of preprints since the 60’s
  - long OA tradition
  - >90% of HEP journal articles on arXiv

- dominance of community based information systems
  - arXiv
  - SPIRES
Dominance of community services

SPIRES (1974-)

- network of databases
  - HEP literature, conferences, institutions, experiments, hepnames, jobs
- SLAC – DESY – Fermilab Collaboration
- SPIRES-HEP
  - Metadata for 850k objects, ~800 new records per week
  - Preprints, journal articles, conference contributions, books, grey literature
  - since 1974, web server since 1991
  - 100k searches/day
- high data quality, manually curated, comprehensive coverage
- high acceptance, user involvement

But:
- outdated technology from the 70’s
Invenio (2002-)

- digital multimedia library system
- platform for CERN Document Server (CDS)
- powerful search engine
  - Google-like speed for up to 5M records
  - combined metadata, reference and fulltext search
- flexible metadata (MARCXML, multimedia)
- personalization and collaborative features
- modular architecture
- Apache/Python/MySQL
- GNU General Public Licence
  - ~30 instances worldwide
ingestion
dissemination
INSPIRE development

- 2007: Inception, feasibility study
- 2008: user-level functionalities
  - data conversion
  - citation analysis, search syntax, output formats…
- 2009: cataloguing functionalities
  - metadata maintenance and enrichment tools
- 2010: workflow
  - harvesting, cataloguing…
- April 2010: public beta version
  http://inspirebeta.net
Bibliographic Content

● SPIRES content (plus part of CDS):
  journal articles, conference proceedings, preprints, experimental notes, theses

● going beyond SPIRES:
  conference slides, multimedia, software, high-level research data…

● going back before 1974

● more material from neighboring disciplines
  astrophysics, nuclear physics, mathematics…
  cited by core HEP articles
“Fulltext” repository

- all freely accessible articles
  - esp. “endangered” material
- access restricted articles
  - “hidden archive”
  - first agreements with Springer and APS
- historical material
  - scanning of old preprint series
- beyond articles
  - slides, multimedia, software, wikis…
  - independent citable objects
INSPIRE features I

- Advanced search functionality
  - Google-like freetext search
  - Complex second-order searches

Example:

Find the most influential HEP core papers that cite the Hitchin article „Generalized Calabi-Yau manifolds“ but don’t cite any papers by Polchinski

```text
```
INSPIRE features II

- detailed record pages
  - abstract, keywords, references, citations, fulltext, figures
  - various export formats

- comprehensive author pages
  - affiliation history, coauthors, frequent keywords, article classification, citation summary

- citation analysis
  - cited by, co-cited with, self-citations, citation history

- taxonomy based classification
HEP taxonomy

hierarchical structure of all important

- HEP concepts (dynamical symmetry breaking)

providing

- synonyms (dynamically broken)
- related terms (spontaneous symmetry breaking)
- broader/narrower (symmetry breaking)
- definitions
- subject areas (high-energy physics – theory)
Taxonomy applications

- fast automatic generation of keywords
  - enabling e.g. prompt alerts
  - manually curated afterwards

- automatic selection of HEP relevant articles
  - no longer time delay in border areas due to manual selection

- improved search algorithm (planned)
  - A search for „SUSY“ will also find „supersymmetry“
  - narrow/broaden search

- user tagging (planned)
  - improve Inspire generated classification
  - improve taxonomy
Author identification

- **INSPIRE author id**
  - compatible with other identification schemes
  - active participation in ORCID

- **author disambiguation**
  - using e.g. lab id’s, affiliation history, coauthors and more
  - 22,000 INSPIRE-id’s already assigned

- **automatic association of papers with authors**
  - using info on affiliations, coauthors, research topics, from publishers
    - G. Chen: 963 docs, 21 real authors, only 22 docs not assigned, 97.2% success rate
  - INSPIRE-id part of author lists of large collaborations
Coming sooner…

- **personalization**
  - personal accounts, bookshelves, display formats, e-mail alerts, RSS feeds
  - collaborative tools, user groups
- **claim my papers**
- **user tagging**
- **fulltext search**
  - snippet display
- **plot extraction, figure caption search**
  - captions in TeX, display via jsMath, TeX symbols searchable
- **user submission**
  - paper centric (articles, supplementary material) and beyond
… or later

- innovative metrics
- semantic analysis
- content indexing of plots and tables
- recommender systems
  - combining citations, keywords, fulltext, usage pattern data...
- open API for 3rd party tools and searching
- object aggregation (OAI-ORE)
- OAIS standards for long-term document preservation

inspire
Partnerships

- researchers
  - user tagging, user submission
  - improved correction interfaces
  - feedback driving future developments

- information providers
  - close alliance with arXiv
  - data exchange with publishers/databases
  - standardized author identities

- neighboring fields
  - open harvesting and searching
  - ADS (SAO/NASA Astrophysics Data System)
  - DML