Annette Holtkamp; Salvatore Mele; Tibor Šimko; Tim Smith

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

In: Petr Sojka (ed.): Towards a Digital Mathematics Library. Paris, France, July 7-8th, 2010. Masaryk University Press, Brno, Czech Republic, 2010. pp. 83--92.

Persistent URL: `http://dml.cz/dmlcz/702577`

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

Annette Holtkamp*, Salvatore Mele, Tibor Šimko, and Tim Smith
on behalf of the INSPIRE Collaboration

CERN, 1211 Geneve 23, Switzerland
`Annette.Holtkamp@cern.ch` `Salvatore.Mele@cern.ch`
`Tibor.Simko@cern.ch` `Tim.Smith@cern.ch`
`http://cern.ch`

Abstract. High-Energy Physics (HEP) has a long tradition in pioneering infrastructures for scholarly communication, and four leading laboratories are now rolling-out the next-generation digital library for the field: INSPIRE. This is an evolution of the extraordinarily successful, 40-years old SPIRES database. Based on the Invenio software, INSPIRE already provides seamless access to almost 1 million records, which will be expanded to cover multimedia, data, software, wikis. Services offered include citation analysis, fulltext search, extraction of figures from fulltext and search in figure captions, automatic keyword assignment, metadata harvesting, retrodigitization, ingestion and automatic display of LaTeX, and storage of supplementary materials like Mathematica notebooks. New services are in different phases of design or implementation, in strategic partnerships with all other information providers in the field and neighbouring disciplines, including; automatic author disambiguation, user tagging, crowdsourcing of metadata curation, automatic document classification, semantic analysis, innovative metrics, recommender systems, object aggregation with OAI-ORE definition, integration of OAIS standards for long-term document preservation.

Key words: digital library, high-energy physics, INSPIRE, Invenio, metadata curation

1 Introduction

High-Energy Physics (HEP) takes pride in a long tradition of pioneering infrastructures for scholarly communication, with half a century of practice in preprint dissemination and two decades of expertise in running repositories [1]. It is rapidly evolving its scholarly communication platforms to realise the hopes of the e-science era. With the recent launch of the INSPIRE system [2], HEP scientists are seeing their dream come true of a digital library encompassing the complete corpus of their scientific output and providing state-of-the art information tools to optimize their research workflow.

* on leave of absence from DESY, Notkestr. 85, D-22607 Hamburg, Germany

Petr Sojka (editor): DML 2010, Towards a Digital Mathematics Library, pp. 83–92.
ISBN 978-80-210-5242-0

Global collaboration is needed to create a platform that satisfies the needs of scholars for easy and unrestricted access to comprehensive scientific information in their field and neighboring disciplines and for powerful discovery tools. Thus the four leading HEP laboratories in Europe and the US have joined forces to develop the next-generation information platform, INSPIRE, tailored to the specific needs of the HEP community. CERN, DESY, Fermilab and SLAC are working in synergy with arXiv.org [3], publishers and other information providers in the field to build and operate INSPIRE as an evolution of the extraordinarily successful SPIRES database [4]. Based on the Invenio [5] Open Source digital library software developed at CERN, INSPIRE provides seamless access to almost 1 million records and will in the near future extend its scope to include supplementary material, multimedia, data, software, wikis. It will enable novel text- and data-mining applications and deploy new metrics to assess the impact of articles and authors.

This paper will outline the services currently offered by INSPIRE as well as new features presently being designed and implemented. Since the last decades have witnessed a growth of interdisciplinary ties between HEP and mathematics, the focus will be on describing strategies to solve current challenges common to HEP and mathematics.

2 The HEP Information Landscape

HEP scientists work in a relatively small, closely-knit community consisting of 20–30,000 researchers. About 50% of them are theorists writing 80% of all HEP papers in small global collaborations of less than 10 authors. The other half are experimental physicists mostly working at big research centres in large global collaborations, exemplified by the fact that the recent papers published by the LHC collaborations at CERN carry more than 2,000 authors.

Particle physicists have always been driven by the need for rapid sharing of ideas and research results. This desire for speed in combination with the global interconnectivity of the HEP research community led to the early development of a preprint culture. Today, more than 90% of all HEP journal articles are submitted to arXiv.org. But already in the 1960s it was common practice for HEP authors and institutes to distribute paper copies of articles worldwide before their publication in journals. In 1974, out of a library catalog of these preprints, the SPIRES-HEP database was born [6]. In Dec 1991, SPIRES-HEP became the first database on the web. Some months before, the first e-print archive, now known as arXiv.org, was set up. Since then, a symbiotic relationship has developed between these two community-driven information systems.

The SPIRES database, jointly run by SLAC, DESY and Fermilab, now contains more than 850k bibliographic records (preprints, journal articles, conference contributions) covering the entire HEP literature and many papers from related fields. Its human-curated metadata includes links to fulltext, author affiliations, citations, publication information, keywords from a HEP taxonomy and much more. Currently, about 100k searches are performed per day.

As a consequence of the decades worth of trusted, curated content it contains and its user-driven evolution SPIRES enjoys an overwhelming popularity within its worldwide user community. In a survey performed in the spring of 2007, 91.4% of the participants mentioned the community-based systems SPIRES and arXiv as their favourite information source [7]. The poll also highlighted the fact that SPIRES' aging technological infrastructure presented a severe obstacle to fulfilling the future information needs of its user community. Therefore in May 2007, at the 1st HEP/PPA Information Resource Summit [8] the SPIRES collaboration joined forces with CERN to develop INSPIRE, the next-generation gateway to all HEP relevant information. A public beta version is accessible since April 2010 [2].

3 INSPIRE Overview

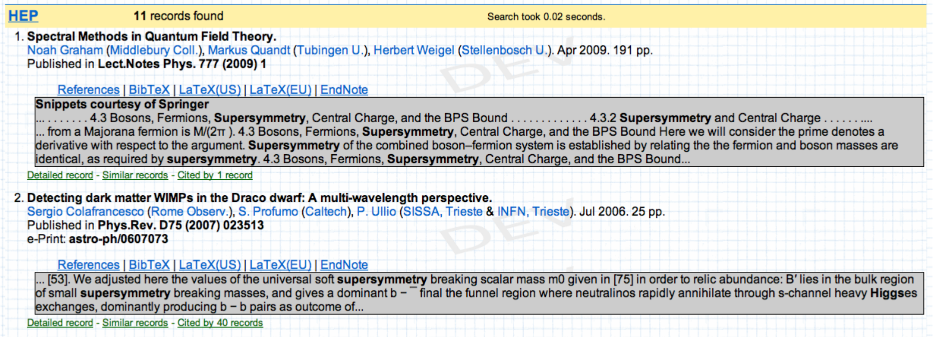
By migrating SPIRES to the Invenio platform, a modern open-source multimedia digital library software developed at CERN, cutting-edge information tools have been put at the disposal of particle physicists. Invenio's strengths include speed, scalability to millions of records, a flexible metadata model supporting a variety of document types (articles, photos, videos), personalization and collaborative features, and a multilingual interface with support for 25 languages. Invenio uses a modular architecture and relies on acknowledged standards like MARCXML [9] for storing bibliographic data or OAI-PMH for metadata exchange [10]. As part of the Open Source community, the software is available under the GNU General Public License, and has over 25 production instances worldwide.

Besides supporting the traditional SPIRES specific search syntax, INSPIRE enables Google-like free keyword searches across metadata and fulltext. Invenio's powerful search engine allows most queries to be executed in a fraction of a second, even for a repository with a million records.

Moving beyond SPIRES' traditional role as a metadata store, INSPIRE will act as a fulltext repository hosting all freely accessible preprints, journal articles, conference contributions, and theses, enabling fulltext search and displaying snippets of text surrounding search terms on the results page, as shown in Fig. 1. Negotiations with publishers are under way to extend this functionality to access-restricted articles, especially with a view to articles predating arXiv. A first agreement has been signed with Springer in April 2010.

For each article, a detailed page shows abstract, keywords, publication information, links to different fulltext versions and to a wealth of additional information. Work is in progress to extract figures from all arXiv papers recorded in INSPIRE and to display them as a film strip on the detailed record page, as exemplified in Fig. 2 on the following page.

The figures are extracted from arXiv source tarballs and associated to paper records. The TEX sources of arXiv papers are parsed in order to extract the captions associated with each figure. The TEX formatting used in captions is stored as such in bibliographic records and is displayed in the user browser



HEP 11 records found Search took 0.02 seconds.

1. **Spectral Methods in Quantum Field Theory.**
Noah Graham (Middlebury Coll.), Markus Quandt (Tubingen U.), Herbert Weigel (Stellenbosch U.). Apr 2009. 191 pp.
Published in **Lect.Notes Phys. 777 (2009) 1**

References | BibTeX | LaTeX(US) | LaTeX(EU) | EndNote

Snippets courtesy of Springer
... . 4.3 Bosons, Fermions, **Supersymmetry**, Central Charge, and the BPS Bound . . . . 4.3.2 **Supersymmetry** and Central Charge . . . .
... from a Majorana fermion is M/(2π). 4.3 Bosons, Fermions, **Supersymmetry**, Central Charge, and the BPS Bound Here we will consider the prime denotes a derivative with respect to the argument. **Supersymmetry** of the combined boson–fermion system is established by relating the the fermion and boson masses are identical, as required by **supersymmetry**. 4.3 Bosons, Fermions, **Supersymmetry**, Central Charge, and the BPS Bound...

Detailed record - Similar records - Cited by 1 record

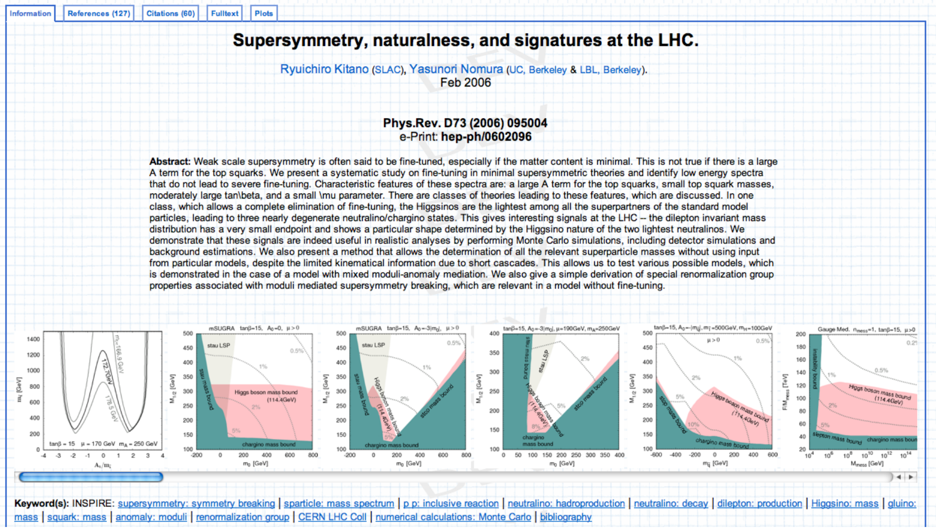
2. **Detecting dark matter WIMPs in the Draco dwarf: A multi-wavelength perspective.**
Sergio Colafrancesco (Rome Observ.), S. Profumo (Caltech), P. Ullio (SISSA, Trieste & INFN, Trieste). Jul 2006. 25 pp.
Published in **Phys.Rev. D75 (2007) 023513**
e-Print: **astro-ph/0607073**

References | BibTeX | LaTeX(US) | LaTeX(EU) | EndNote

... [53]. We adjusted here the values of the universal soft **supersymmetry** breaking scalar mass m0 given in [75] in order to relic abundance: B′ lies in the bulk region of small **supersymmetry** breaking masses, and gives a dominant b − ¯ final the funnel region where neutralinos rapidly annihilate through s-channel heavy **Higgs**es exchanges, dominantly producing b − b pairs as outcome of...

Detailed record - Similar records - Cited by 40 records

Fig. 1. Search results page with fulltext snippets



Supersymmetry, naturalness, and signatures at the LHC.

Ryuichiro Kitano (SLAC), Yasunori Nomura (UC, Berkeley & LBL, Berkeley).
Feb 2006

Phys.Rev. D73 (2006) 095004
e-Print: **hep-ph/0602096**

Abstract: Weak scale supersymmetry is often said to be fine-tuned, especially if the matter content is minimal. This is not true if there is a large A term for the top squarks. We present a systematic study on fine-tuning in minimal supersymmetric theories and identify low energy spectra that do not lead to severe fine-tuning. Characteristic features of these spectra are: a large A term for the top squarks, small top squark masses, moderately large tan\beta, and a small \mu parameter. There are classes of theories leading to these features, which are discussed. In one class, which allows a complete elimination of fine-tuning, the Higgsinos are the lightest among all the superpartners of the standard model particles, leading to three nearly degenerate neutralino/chargino states. This gives interesting signals at the LHC -- the dilepton invariant mass distribution has a very small endpoint and shows a particular shape determined by the Higgsino nature of the two lightest neutralinos. We demonstrate that these signals are indeed useful in realistic analyses by performing Monte Carlo simulations, including detector simulations and background estimations. We also present a method that allows the determination of all the relevant superparticle masses without using input from particular models, despite the limited kinematical information due to short cascades. This allows us to test various possible models, which is demonstrated in the case of a model with mixed moduli-anomaly mediation. We also give a simple derivation of special renormalization group properties associated with moduli mediated supersymmetry breaking, which are relevant in a model without fine-tuning.

tanβ = 15 μ = 170 GeV m_A = 250 GeV
mSUGRA tanβ=15, A₀=0, μ>0
stau LSP
Higgs boson mass bound (114.4GeV)
chargino mass bound
mSUGRA tanβ=15, A₀=−3|m₀|, μ>0
stop mass bound
tanβ=15, A₀=−3|m₀|, μ=190GeV, m_A=250GeV
tanβ=15, A₀=−|m_q̃|, m_l̃ =500GeV, m_H=100GeV
Gauge Med. n_mess=1, tanβ=15, μ>0
slepton mass bound
m₀ [GeV]
M_mess [GeV]

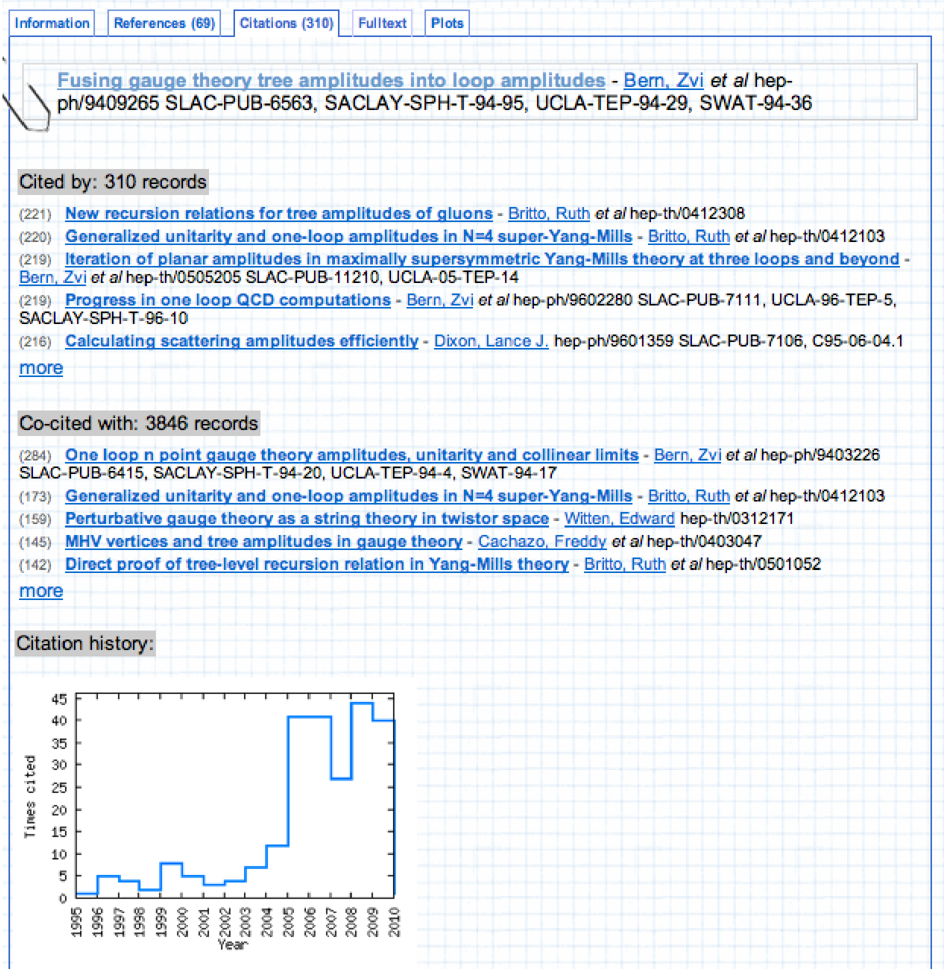
Keyword(s): INSPIRE: supersymmetry: symmetry breaking | sparticle: mass spectrum | p p: inclusive reaction | neutralino: hadroproduction | neutralino: decay | dilepton: production | Higgsino: mass | gluino: mass | squark: mass | anomaly: moduli | renormalization group | CERN LHC Coll | numerical calculations: Monte Carlo | bibliography

Fig. 2. Detailed article page with plot slider

via the jsMath library [11]. Storing of captions in TEX permits them to be independently searchable, not only for words and phrases, but for TEX symbols as well.

A further strength of INSPIRE is its citation analysis. The "co-cited with" network gathers information about papers which are frequently cited together with the paper of interest, opening new paths to find related articles. A citation history graph visualizes citation counts of an article over time, enabling easy discovery of various characteristic citation time patterns such as a "sleeping beauty", one example being shown in Fig. 3 on the next page.

The metadata, full-text, figure caption, citation, and other search indexes can be mutually combined, contributing to the unprecedented level of



Information | References (69) | Citations (310) | Fulltext | Plots

Fusing gauge theory tree amplitudes into loop amplitudes - Bern, Zvi *et al* hep-ph/9409265 SLAC-PUB-6563, SACLAY-SPH-T-94-95, UCLA-TEP-94-29, SWAT-94-36

Cited by: 310 records

(221) New recursion relations for tree amplitudes of gluons - Britto, Ruth *et al* hep-th/0412308
(220) Generalized unitarity and one-loop amplitudes in N=4 super-Yang-Mills - Britto, Ruth *et al* hep-th/0412103
(219) Iteration of planar amplitudes in maximally supersymmetric Yang-Mills theory at three loops and beyond - Bern, Zvi *et al* hep-th/0505205 SLAC-PUB-11210, UCLA-05-TEP-14
(219) Progress in one loop QCD computations - Bern, Zvi *et al* hep-ph/9602280 SLAC-PUB-7111, UCLA-96-TEP-5, SACLAY-SPH-T-96-10
(216) Calculating scattering amplitudes efficiently - Dixon, Lance J. hep-ph/9601359 SLAC-PUB-7106, C95-06-04.1
more

Co-cited with: 3846 records

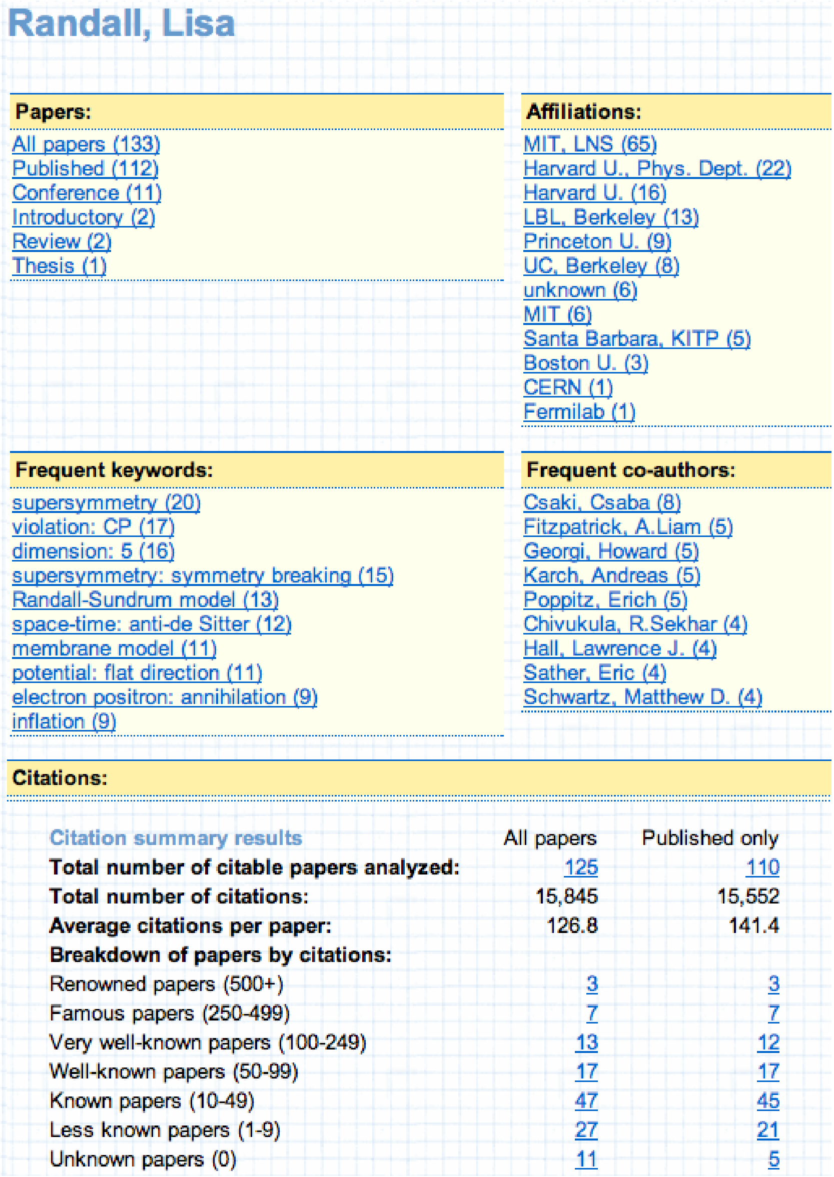
(284) One loop n point gauge theory amplitudes, unitarity and collinear limits - Bern, Zvi *et al* hep-ph/9403226 SLAC-PUB-6415, SACLAY-SPH-T-94-20, UCLA-TEP-94-4, SWAT-94-17
(173) Generalized unitarity and one-loop amplitudes in N=4 super-Yang-Mills - Britto, Ruth *et al* hep-th/0412103
(159) Perturbative gauge theory as a string theory in twistor space - Witten, Edward hep-th/0312171
(145) MHV vertices and tree amplitudes in gauge theory - Cachazo, Freddy *et al* hep-th/0403047
(142) Direct proof of tree-level recursion relation in Yang-Mills theory - Britto, Ruth *et al* hep-th/0501052
more

Citation history:

Fig. 3. Citation page with co-citations and citation history graph

search flexibility INSPIRE will offer. For example, the search "author:Ellis caption:model cited:10→20 reference:astro" will return all papers by an author named "Ellis" that contain the word "model" in a figure caption, have been cited between 10 and 20 times, and that reference some astrophysical arXiv paper.

Another example of INSPIRE's novel features are author pages which are built dynamically, as exemplified in Fig. 4 on the following page. An author page provides a comprehensive profile of a scientist, containing information on affiliation history, research subjects, frequent coauthors, breakdown of articles according to their type (journal article, conference contribution, lectures etc) as well as breakdown of articles with respect to their citation counts. The "citation summary” format is suited to give some indication of the impact not only of a single scientist, but may also be applied to institutions, countries or the output of any query.



Randall, Lisa

Papers:

All papers (133)
Published (112)
Conference (11)
Introductory (2)
Review (2)
Thesis (1)

Affiliations:

MIT, LNS (65)
Harvard U., Phys. Dept. (22)
Harvard U. (16)
LBL, Berkeley (13)
Princeton U. (9)
UC, Berkeley (8)
unknown (6)
MIT (6)
Santa Barbara, KITP (5)
Boston U. (3)
CERN (1)
Fermilab (1)

Frequent keywords:

supersymmetry (20)
violation: CP (17)
dimension: 5 (16)
supersymmetry: symmetry breaking (15)
Randall-Sundrum model (13)
space-time: anti-de Sitter (12)
membrane model (11)
potential: flat direction (11)
electron positron: annihilation (9)
inflation (9)

Frequent co-authors:

Csaki, Csaba (8)
Fitzpatrick, A.Liam (5)
Georgi, Howard (5)
Karch, Andreas (5)
Poppitz, Erich (5)
Chivukula, R.Sekhar (4)
Hall, Lawrence J. (4)
Sather, Eric (4)
Schwartz, Matthew D. (4)

Citations:

Citation summary results	All papers	Published only
Total number of citable papers analyzed:	125	110
Total number of citations:	15,845	15,552
Average citations per paper:	126.8	141.4
Breakdown of papers by citations:		
Renowned papers (500+)	3	3
Famous papers (250-499)	7	7
Very well-known papers (100-249)	13	12
Well-known papers (50-99)	17	17
Known papers (10-49)	47	45
Less known papers (1-9)	27	21
Unknown papers (0)	11	5

Fig. 4. Author page aggregating various information

As a clear response to a request from the community whose experimental collaborations now count author lists of over 2,500 scientists, work is under way to uniquely identify authors and link them unambiguously to their scientific output. INSPIRE has developed its own author identification scheme and is a leading participant in the ORCID initiative [12] to establish interoperability between different author identification projects and resolve the problem of author ambiguity on a global scale. Based on its detailed knowledge about a scientist's research topics, coauthor network, affiliation history, citation patterns and so on, INSPIRE is able to resolve author name ambiguities and to calculate degrees of probability for an article to be written by a certain author. To give some indication of the performance, for a set of 963 documents with author name written as "Chen, G", 21 distinct real authors have been identified. Only 22 out of 963 documents were not associated with one of these authors, giving the algorithm in this case a success rate of 97.2%. As a next step, an interface is under development to allow registered authors to claim their papers, further feeding into the overall data quality for that given author and, through the co-authorship network, of the whole database. Articles are categorized by probability of ownership and displayed to the presumable author who is asked to confirm or reject these attributions. In addition, an option is offered to claim papers that have not been suggested or to submit papers not yet included in INSPIRE. The author names are internally represented in Unicode UTF-8 character encoding within INSPIRE, enabling association of translated author names with the names in their original languages, including ideograms.

4 Outlook

The beta version of INSPIRE is now operational, reproducing and improving the basic services which have powered the community, with SPIRES, over decades:

- central access to the complete HEP literature
- high-quality human-curated metadata
- very fast search engine enabling Google-like free keyword searches
- taxonomy-based classification
- comprehensive author pages
- extensive citation analysis

The next important step will be to roll-out personal accounts. These will be activated within the next few months, enabling features like personal bookshelves, email notification alerts and RSS feeds, personalized display formats and tools for sharing information within a collaboration. Powerful incentives for the creation of personal accounts will be the claiming of articles, an improved system for notifying missing references, and tools for annotating and organising bibliographies. These services are known from SPIRES to have been on the "desiderata" list of the community for a long time.

Personal accounts will also open the door to the porting of tested Web 2.0 models of user-generated content into a large-scale digital library. In the 2007 user poll, 63% of the respondents expressed their willingness to spend at least half an hour per week on enriching the database content [7]. A first attempt to harness this amazing potential will be to encourage users to tag content.

An important evolution of INSPIRE with respect to SPIRES is the possibility of hosting documents and other materials, rather than just linking to them. Users registered in Inspire will therefore have an opportunity to upload written material that they would not submit to the arXiv. A classic example is older material, from theses to unpublished documents, that they would like to see online, but, not being of recent origin, they do not want included in arXiv alerts. As an immediate extension of this possibility Inspire can allow, within moderate storage limits, the uploads of other kind of documents, like Mathematica notebooks, software source code, additional graphs, small data sets—not only as supplementary material directly attached to articles but, moving beyond the article-centric model, as independent citable objects. The centralisation of this material, away from personal web pages, and in a clear format linked to publications, is another long-standing desire of the community. A corollary of this rich harvest of additional objects will be their aggregation (either from the curators, or crowdsourced) into a single view of the same idea. The OAI-ORE standards definitions [13] are under consideration as a scheme to aggregate related objects.

Semantic techniques for information classification retrieval are currently under development, based on a taxonomy of HEP concepts [14]. By exploiting synonyms, more comprehensive search results will be achieved. Another application currently being refined is the automatic categorization of material on ingestion so that a paper is automatically recognized e.g. as a conference talk on renormalization in perturbative quantum field theory or as a thesis on the electroweak model in noncommutative geometry. Other features to come are faceting of search results and a recommender system to suggest similar material based on combined citations, keywords, and usage pattern data.

Thanks to its role as central HEP information system, INSPIRE is ideally placed to become an essential agent in digital preservation of particular classes of documents. On the grey literature side, a lot of effort has already been invested in retrodigitizing research papers and theses of the four laboratories running INSPIRE. These were inaccessible so far and are now archived in a persistent digital format. Services of preservation on demand for users will be made possible for all additional material discussed above, from small data files to Mathematica notebooks, from conference slides to multimedia. An additional incentive for preservation will be the fact that INSPIRE will make this material discoverable and citable. Another use case for preservation is the documentation that large experimental collaborations produce in support of their scientific analyses which is today locked either in notes or in twikis. These are as persistent as the organizations which created them, poised to move on to other scientific endeavours. An effort is under way for the ingestion of this material in INSPIRE, linked to the original publication to which it refers, with correct provenance information and access rights reflecting the policies of the scientific groups which prepared this material. To this extent necessary steps will be taken to make INSPIRE OAIS compliant [15]. As an aside, this process will also enable innovative metrics to take into account nontraditional forms of scientific results.

HEP as a field has long been vigilant to seize interdisciplinary opportunities. A notable example is the large overlap in literature and in scientists with astronomy, astrophysics and astroparticle physics. As a consequence, the digital libraries of these fields are moving closer together. Astronomy and astrophysics have long relied on the ADS (Astrophysics Data System) [16] run by the Harvard Smithsonian Astronomy Observatory under a NASA grant. Starting with a rich metadata exchange the collaboration between ADS and INSPIRE is evolving towards a joint curation of records of common interest as well as a joint development of full-text search and recommender systems. This will be facilitated by the move of part of the ADS operations to the Invenio platform as well.

It is easy to imagine that a similar, tight, collaboration could be initiated with an emerging digital library for mathematics. A large amount of mathematical tools are used by theoretical HEP scientists, which could benefit from a more powerful set of discovery and retrieval opportunities through the interfacing of INSPIRE and such a digital library for mathematics. At the same time, cross-disciplinary records could be curated only once, information on author identification across the systems could be streamlined, and citations could be followed seamlessly.

In conclusion, several lessons have been learnt in the inception of INSPIRE, the transition from SPIRES to INSPIRE and the planning of future services. The most relevant are that a careful analysis of users' needs and desires should be the driving force of all planning, and that a wide-range search for synergies and agreements across all information providers can accelerate development and deployment of new services. Both lessons may seem obvious, but there is always a risk in these kind of projects that user-pull loses against technology-push, collaboration against silo mentality.

The realisation of a large, federated, interoperable e-infrastructure for scholarly communication is coming closer and closer, and neighbouring fields have a unique opportunity to move together to deliver key services to their scientific communities.

References

1. R. Heuer, A. Holtkamp and S. Mele, Innovation in Scholarly Communication: Vision and Projects from High-Energy Physics *Info. Ser. and Use* 28 (2008) pp. 83–96, arXiv:0805.2739v1, doi:10.3233/ISU-2008-0570
2. http://inspirebeta.net
3. http://arXiv.org
4. http://www.slac.stanford.edu/spires/
5. http://invenio-software.org/
6. L. Addis http://www.slac.stanford.edu/spires/papers/history.html

7. A. Gentil-Beccot et al., Information Resources in High-Energy Physics: Surveying the Present Landscape and Charting the Future Course, *J. Am. Soc. Inf. Sci. Technol.* 60 (2009) pp. 150–160, arXiv:0804.2701v2, doi:10.1002/asi.20944
8. http://indico.cern.ch/event/11611
9. http://www.loc.gov/standards/marcxml
10. http://www.openarchives.org/OAI/openarchivesprotocol.html
11. http://www.math.union.edu/~dpvc/jsMath/
12. http://www.orcid.org
13. http://www.openarchives.org/ore/
14. http://www-library.desy.de/akw/HEPont.rdf
15. Reference Model for an Open Archival Information System (OAIS), http://public.ccsds.org/publications/archive/650x0b1.pdf
16. http://adswww.harvard.edu/