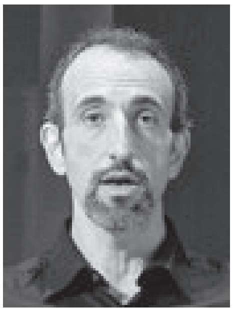
Sünje Dallmeier-Tiessen, Salvatore Mele

Integrating Data in the Scholarly Record: Community-Driven Digital Libraries in High-Energy Physics

Foto: Lou Woodley

Sünje Dallmeier-Tiessen

Foto: Jason Wen



Salvatore Mele

This brief article describes how research data can be integrated into a large-scale digital library. INSPIRE serves a community where data sharing is emerging. New open science services are developed to support the grass root interest of the community. Digital library staff collaborates closely with researchers to design new services. The services not only focus on a technical integration of data in INSPIRE, but also enable open research data to become a central part of scholarly communication in the field. Research data becomes discoverable, reusable and citable. We present the very first examples of data citation and tracking in the field.

Der Beitrag beschreibt die Integration von Forschungsdaten in eine umfangreiche Digitale Bibliothek. INSPIRE bedient die gesamte Fachgemeinschaft der Hochenergiephysik, welche in der Vergangenheit relativ wenig Erfahrung im offenen Umgang mit Forschungsdaten hatte. Parallel zu dem stetig wachsenden Interesse werden nun Services als Unterstützung entwickelt. Zur Spezifizierung der Services arbeiten die MitarbeiterInnen der Digitalen Bibliothek sehr eng mit den Wissenschaftlern zusammen. Die Services fokussieren nicht nur eine technische Integration der Forschungsdaten, sondern zielen darauf ab, Forschungsdaten als nachnutzbaren und zitierbaren wichtigen Teil der Wissenschaftskommunikation zu nutzen. Vorgestellt werden erste Beispiele der Datenzitation und deren Tracking in INSPIRE.

Introduction

Open Access

»Data is the new oil«, says Neelie Kroes, Vice President of the European Commission responsible for the Digital Agenda [Kroes, 2012]. The contemporary technological revolution has made it possible to process an unprecedented amount of data, refine hypotheses, and, most impressively, share this information in real time with the entire scientific community, and the civil society. This opportunity underpins a deep discussion on the ethics of Open Science and the services that become necessary to publish, retrieve, preserve such open data. Research in the past years has shown that disciplinary challenges and practices are different, especially in sectors where personal data might be part of the research process, e. g. in the Humanities and Social Science or the Life Science [Schaefer et al., 2012]. At the same time a strong commitment to sharing data has evolved in other fields, dating back to the sequencing of the human genome [Wellcome Trust, 2012] and now expanding as new generations of scientists join the ranks of academia. It is by now a fact that research data can be integrated into scholarly communication, independent of disciplinary challenges. A fact that has transformed libraries forever.

This short article describes an example of integration of research data in a digital library, called INSPIRE [http://inspirehep.net]. It is an example from the High-Energy Physics (HEP) community, which epitomizes »big data« science. HEP has a closely knit international community, it concentrates experimental studies in a few large scale laboratories, such as CERN, while theorists in universities around the world suggest hypotheses and, eventually, make sense of the experimental data.

The digital library INSPIRE

For almost five decades, the HEP community has relied on a centralized bibliographic database, which dates back to the late 1960's: SPIRES, the Stanford Physics Information Retrieval System. It grew to become its discipline-wide digital library and it recently morphed into a new service, INSPIRE [Gentil-Beccot, A. et al, 2008]. This digital library is based on the CERN Invenio Open Source digital library software and, in the tradition of the field, is run as a collaboration of five laboratories in Asia, Europe and the United States: CERN (European Organization for Nuclear Research, Switzerland), DESY (Deutsches Elektronen-Synchrotron, Germany), Fermilab (Fermi National Accelerator Laboratory, United States), IHEP (Institute of High Energy Physics, China) and SLAC (National Accelerator Laboratory, United States). INSPIRE serves as a content aggregator of over a million preprints, published content, conference proceedings, as well as theses and other disciplinary-specific examples of grey literature, such as the working papers that large collaborations prepare in advance of conferences, or unpublished supplementary material and illustrations. INSPIRE offers access to full-text material from many different sources. It also indexes over half a million Open Access full-text works directly, alongside several hundreds of thousands of articles provided for indexing by key publishers in the field. INSPIRE runs its own author and contributor disambiguation services, based on an automated algorithm combined with user input, soon to be fully integrated with ORCID (Open Researcher and Contributor ID). Those assets make INSPIRE the central platform used to access relevant HEP literature, with more than two searches per second, around the clock. This central role engages the community to contribute back: INSPIRE receives several thousands of user requests a year for metadata enrichment and for the addition of relevant content.

Research data in HEP

HEP aims to understand the building blocks of matter and how these are kept together, by analyzing increasingly rare processes. These studies require vast amounts of data to obtain evidence of otherwise elusive processes. The LHC (Large Hadron Collider) project in its first years has already produced 100PB of data, including a complex simulation needed for the understanding of the experimental apparatuses. This forms one end of the data spectrum. Management of this data requires complex global tiered e-infrastructures, such as the worldwide LHC computing grid (WLCG). At the other end of the data spectrum are highly processed sets of data, essentially underpinning tables and figures in the final publication of results [Dallmeier-Tiessen, 2013]. The Data Preservation in HEP (DPHEP) study group has articulated the different levels in which data can exist, and the challenges of preserving, opening and retrieving each of them [Data Preservation in High-Energy Physics, 2012].

On the »raw data« end of the spectrum, access to the WLCG is meant for researchers participating in the large consortia producing and analyzing data, which can count up to 3.000 scientists. Conversely, there is an increasing awareness of the opportunities of widening access to data on the publication end of the spectrum. This process has roots in a data repository named HEPData [http://hepdata.cedar.ac.uk/] which has been used in the field for over 20 years to allow theorists to re-use numerical versions of data often only published by experimentalists in a graphical form. Through dedicated submission workflow, experimental collaborations provide these data, underpinning publication. These are indexed with their relation to the published article (as indexed by SPIRES, the predecessor of INSPIRE). A »data abstract« is prepared as part of the curation process. Over the last couple of years, this model has expanded and experimental collaborations have experimented with increased sharing of research material. This drive has emerged from the long tradition of Open Access to publications in HEP, as funding agencies became increasingly interested in Open Data. The next two sections describe this process and the challenges and opportunities it posed to rethink the role of (community-driven) digital libraries to underpin Open Data.

Integrating research data (and more) into the digital library INSPIRE

As the central platform for information discovery in HEP, INSPIRE has been naturally involved from the onset in discussions on data preservation and access. This has allowed to understand a demand for new data related services in its early stages and in synergy with the community that this digital library serves. Perhaps unsurprisingly, the call for the community was for the digital library to serve a similar role for data as it does for »traditional« text publications: as an aggregator. Exactly in the same way INSPIRE presents grey-literature, pre-prints, published articles and conference proceedings on a subject, both served from the site or linked to the outside version of records, the community not only expected it to point to data, and its relation to articles, existing elsewhere, but to act as a ›safe harbor‹ of last resort for direct submission of some materials as well. Interestingly, these conversations happened while a suite of third-party community or commercial services have emerged in the open data landscape (e. g. Dataverse [http://thedata.org], Figshare [http://figshare.com] and Zenodo [http://zenodo.org]).

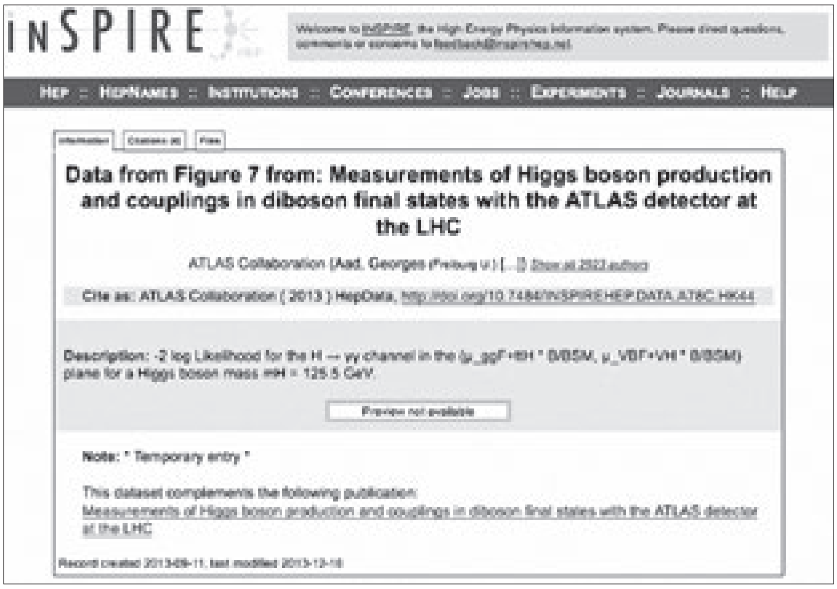
data spectrum

As a consequence, INSPIRE has been and is being set up to serve such a role, expanding its expertise with text materials. This expertise is crucial when it comes to metadata, as the same cataloguing standards as for core literature of the field are now applied throughout the holdings, from hosted data to the one existing elsewhere. This is indispensible to allow users to discover and retrieve materials without much change in their browsing and searching habits. This approach can, probably, be exemplary for other digital libraries expanding in the data realm. The integration of research data into INSPIRE had an exploratory nature at first. It had to be understood, which data could be integrated, how, and in particular, how the data and their metadata would be displayed. We report on a few examples of this process.

For a start, the community HEPData repository was integrated more closely with INSPIRE. Discoverability was improved with additional data tabs appearing alongside the records of existing publications. Most importantly, each data object exists as an individual record in the INSPIRE core collection, treated as »first class« objects, in parallel to the text publications. This shift from paper-centric paradigm used over four decades, to a research-object-centric approach allows to leverage all other services available in the software stack of INSPIRE. These individual data records are indexed and searchable for all metadata fields. They are attributed to disambiguated authors. They also get a persistent identifier assigned, a DOI, and become citable object (see Fig. for an example). Notably, this approach has altered the scientific discourse in the field, with important data, underpinning the discovery of the Higgs boson now made citable, and effectively

research-object-centric approach

Credit: inspirehep.net


INSPIRE
Welcome to INSPIRE, the High Energy Physics information system. Please direct questions, comments or concerns to feedback@inspirehep.net
HEP :: HEPNAMES :: INSTITUTIONS :: CONFERENCES :: JOBS :: EXPERIMENTS :: JOURNALS :: HELP
Information
Citations (4)
Files
Preview not available

Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC

ATLAS Collaboration (Aad, Georges (Freiburg U.)) [...] Show all 2922 authors

Cite as: ATLAS Collaboration (2013) HepData, http://doi.org/10.7484/INSPIREHEP.DATA.A78C.HK44

Description: -2 log Likelihood for the H → γγ channel in the (μ_ggF+ttH * B/BSM, μ_VBF+VH * B/BSM) plane for a Higgs boson mass mH = 125.5 GeV.

Note: * Temporary entry *

This dataset complements the following publication:
Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC

Record created 2013-09-11, last modified 2013-12-16

Fig.: A dataset is available on INSPIRE. It is displayed with its corresponding DOI, minted by the digital library. The tab »citations« indicates that the dataset has been published four times, and allows to navigate to the cited articles, exploring the corresponding network

cited by third-party theoretical studies (see the »citations« tab in Fig.) [Aad, G., 2013].

Another community-driven example, within the HEP ecosystem, has been a direct data upload to INSPIRE. The process has allowed to design submissions workflows, while decisions to handle and display the holding have been incremental, leveraging the work for the previous example of ingestion of the data repository. Again, data is presented alongside the publications, whereas the stand-alone data record is now ingested and curated in INSPIRE, creating a basic set of metadata to allow discovery and re-use [Abrahamyan, S., 2011].

close collaboration with the community

While this is an example of a supplementary material to a publication, whereby the digital library has acted as a ›data publisher‹, each piece of infrastructure and the workflows which have been designed, allow the option of INSPIRE to ingest and publish »standalone« datasets as well. The digital library is therefore ready to address community requests for such data releases in the future.

At the same time, beyond the HEP borders, new services have emerged to cater to the data services demand from a wide open science community, notably Dataverse and Figshare. Members of the HEP community experimented with both platforms and contacted INSPIRE to aggregate these external datasets into the scholarly record accessible to the HEP community. Leveraging the previous lesson of presenting supplementary materials, the workflows were developed to create records in INSPIRE with metadata from those platforms, leveraging the persistent identifiers from the respective platforms. In the case of the Dataverse example, these were a handle [Cranmer et al.] and the DOI standards for Figshare [Cranmer and Kreiss, 2014].

A final example has been a synergy between INSPIRE, the Zenodo data repository and Github, the world-leading code-sharing platform [http://github.com] to create a workflow whereby a piece of code used to generate results in a paper indexed by INSPIRE was made citable. In detail: Zenodo would ingest through specially designed API workflows the code from Github and make it citable [CERN, 2014], therefore bringing the code into the scholarly record. As an aggregator, INSPIRE would augment the scientific article underpinned by the code with a ›data‹ tab, ingesting metadata from Zenodo and, again, present all information in a single place for the perusal of its target community. As a result, now data and code can be indexed in INSPIRE and attributed to their authors, becoming accessible from the pages where the entire »scholarly output« of the researcher is presented, and later exposed through other third-party services, such as those building on ORCID. Even more so, data and code are both made citable objects on INSPIRE. Reuse can be tracked, based on the persistent identifiers assigned, so that the researchers have a direct benefit from the integration, which serves as a further incentive for additional material to be made analogously available.

Lessons learnt

A first exploratory phase for research data and code integration in the INSPIRE digital library has been completed, what allows to abstract some key lessons learnt in recent months. The implementation of the services and workflows has been done in close collaboration with the community, which in turn considers the digital library an important part of the e-infrastructure and the service environment. The demand from pioneers in the community for these services was in fact the catalyst for the digital library staff to evolve in a role of data librarians and curators.

A key thread in the action to integrate (open) data in the scholarly discourse has been the responsibility from the digital library to provide a good set of metadata (possibly beyond the present emerging standard DataCite set [http://datacite.org]). Interestingly, rather than a requirement on the library side, this has often been a request from researchers needing support in terms of standardization. While taking these steps towards Open Science, it is important to reflect on the different roles and functions that INSPIRE has taken,

https://doi.org/10.3196/1864295014614574 https://www.inlibra.com/de/agb - Open Access -

as similar challenges await all digital libraries confronting the data challenges. In this case, the team leverages the request for any given technical implementation to understand the emerging trends and needs of researchers starting to share research data, in order to prototype its next services. In turn the INSPIRE team builds on experience from other disciplines and emerging standards to apply them within the HEP community.

A key lesson learnt is that, in front of the diversity and nuances of the first use cases, no »one size fits all« solution is adequate, but multiple solutions must be offered, often integrating and leveraging emerging third-party services which the community serendipitously adopts. Researchers are very much interested in understanding who reuses their data and code. INSPIRE has adopted persistent identifiers to track reuse, and the initial feedback from the community is the potential for a virtuous circle, where open data services which lead to ease of discovery will generate re-use which, if tracked, inspire further data sharing.

track reuse

The way ahead

The scientific community demand for data services is growing. We are witnessing a »Cambrian Explosion« of services, platforms and options, as data reuse and data citation emerge as very prominent topics in scholarly communication. It is important that libraries close to specific communities, with their expertise, take an immediate leading role in this revolution. In practical terms, libraries can re-define their services starting from concrete user-driven demands, while providing users support and platforms to integrate data in the scholarly records with the highest (internationally-agreed) standards for further re-use. We see aggregation and integration of platforms, rather than competition for users and services, as the new mission to connect researchers and information.

References

Aad, Georges; Abajyan, Tatevik; Abbott, Brad; Abdallah, Jalal; Abdel Khalek, Samah; Abdinov, Ovsat; Aben, Rosemarie; Abi, Babak; Abolins, Maris; et al.: Data from Figure 7 from: Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC. 2013. DOI: 10.7484/INSPIRE HEP.DATA.A78C.HK44

Abrahamyan, S. et al.: Data from Figure 3 from: Search for a New Gauge Boson in Electron-Nucleus Fixed-Target Scattering by the APEX Experiment. 2011. DOI: 10.7484/INSPIREHEP.DATA.PH21.L5RG

CERN: Tool developed at CERN makes software citation easier. http://cern.ch/about/updates/2014/03/tool-developed-cern-makes-software-citation-easier [accessed May 15TH, 2014].

Cranmer, Kyle; Allanach, Ben; Lester, Christopher; Weber, Arne: Replication data for: »Natural Priors, CMSSM Fits and LHC Weather Forecasts«, 2013. http://hdl.handle.net/1902.1/21804

Cranmer, Kyle, Kreiss, Sven: Supplementary Material for ›A Novel Approach to Higgs Coupling Measurements‹. 2014. DOI: 10.6084/m9.figshare.888607

Dallmeier-Tiessen, Sünje: Driver and Barriers in Digital Scholarly Communication. 2013. urn:nbn:de:kobv:11-100216213

Gentil-Beccot, Anne; Mele, Salvatore; Holtkamp, Annette; O'Connell, Heath B.; Brooks, Travis C.: Information resources in High-Energy Physics: Surveying the present landscape and charting the future course. In: Journal of the American Society for Information Science and Technology, 60 (2008).

Kroes, Neelie (2012): Digital Agenda and Open Data. http://europa.eu/rapid/press-release_SPEECH-12-149_en.htm [accessed, May 15TH 2014].

Study Group Data Preservation in High-Energy Physics. Status Report of the DPHEP Study Group: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics, 2012. arXiv: 1205.4667

Schäfer, Angela; Dallmeier-Tiessen, Sunje; Pfeiffenberger, Hans; Pampel, Heinz; Tissari, Satu; Darby, Robert; Giaretta, David; Giaretta, Krystina; Gitmans, Kathrin; et al.: Ten Tales of Drivers & Barriers in Data Sharing. 2012. DOI: 10.5281/zenodo.8308

Wellcome Trust. Statement on genomic data release, www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTD002751.htm [accessed, May 15th, 2014].

The Authors

Sünje Dallmeier-Tiessen, Open Science Research Fellow, CERN, CH-1211 Geneva 23, Tel.: +41 22 – 76 62754, E-Mail: sunje.dallmeier-tiessen@cern.ch, orcid.org/0000-0002-6137-2348

Salvatore Mele, Head of Open Access, CERN,
CH-1211 Geneva 23, Tel.: +41 22 – 76 78603,
E-Mail: Salvatore.Mele@cern.ch,
orcid.org/0000-0003-0762-2235