How partnership accelerates Open Science: High-Energy-Physics and INSPIRE, a case study of a complex repository ecosystem

Sünje Dallmeier-Tiessen, Bernard Hecker, Annette Holtkamp, Salvatore Mele, Heath O'Connell, Kirsten Sachs, Tibor Šimko and Thorsten Schwander for the INSPIRE collaboration

Introduction

Public calls, agency mandates and scientist demand for Open Science are by now a reality with different nuances across diverse research communities. A complex "ecosystem" of services and tools, mostly community-driven, will underpin this revolution in science. Repositories stand to accelerate this process, as "openness" evolves beyond text, in lockstep with scholarly communication.

We present a case study of a *global* discipline, High-Energy Physics (HEP), where most of these transitions have already taken place in a "social laboratory" of multiple global information services interlinked in a complex, but successful, ecosystem at the service of scientists. We discuss our first-hand experience, at a technical and organizational level, of leveraging partnership across repositories and with the user community in support of Open Science, along threads relevant to the OR2013 community.

Background: evolution of the repository and service ecosystem

Starting in the early 1960's, HEP researchers adopted a simple solution to share their results faster than the traditional peer-review and journal distribution system: mailing their research, in its initial *preprint,* form to their colleagues, before sending it to a journal (Heuer et al., 2008). The advantages of speed and the sense of community vastly overweighed the risk of spoofing. Research libraries as CERN/Geneva became hubs, started building metadata records of those preprints at par with other material (Goldschmidt-Clermont, 1965), and further contributed in making this research searchable, retrievable and further re-distributed information. Automation in the late 1960s allowed the SLAC/Stanford library to create a computerized system, SPIRES, as a niche of the ecosystem of communication where all unpublished preprints would be recorded. In the early 1970s DESY/Hamburg joined by indexing relevant published literature in SPIRES (O'Connell, 2000). This first component of the ecosystem went online in December 1991 when SPIRES became the first database on the web and the first website outside of Europe.

In the same year, Paul Ginsparg started the arXiv repository for the HEP theory community, who quickly adopted the service (Ginsparg, 2011). arXiv is now ubiquitous in HEP, as well as in many more fields since. The ecosystem immediately developed a strong synergy in the complementary roles. arXiv became the place where users spontaneously submitted their preprints as the first point of dissemination, while SPIRES added bibiliographic value for users (references, citations, link to published versions) and acted as the entry point for wider literature searches. With the existence of a single entry point, it was a short step to another specialized function needed in the ecosystem: integrating publications and a directory of practitioners in the field, called HEPNames, mostly managed by Fermilab/Chicago.

The ecosystem today: more complexity, connections, more services

The HEP repository ecosystem has evolved fast: CERN has joined and SPIRES has been superseded by INSPIRE (www.inspirehep.net), built on the Invenio Open Source Digital Library software (http://invenio-software.org/). INSPIRE today holds 1 million metadata records, mostly inherited from SPIRES, and augmented by 300'000 full-text documents from arXiv and additional sources. It has an average traffic of over 2 searches per second from the 50'000-strong HEP research community. INSPIRE allows further, complex interconnections in the system:

- INSPIRE is fed from arXiv. Metadata and full-text are indexed for searching, while text-mining extracts authors' name and affiliations, references, figures and figure captions for separate indexing
- INSPIRE crawls the output of publishers of HEP journals, and matches articles to corresponding preprint versions, augmenting their metadata. Win-win agreements allows INSPIRE to index full-text of published versions, increasing discoverability for users and traffic for publishers, through links to their platforms for documents not Open Access.
- INSPIRE feeds arXiv back with information on published versions, augmenting arXiv metadata with DOIs, as well as bibliometrics analysis of citing-cited pairs of articles for further navigation.

Beyond this initial core, other community-based services have independently evolved in this ecosystem. Each specializes in a particular part of the information architecture and flow, around the users' needs, and is integrated with the others:

- The HEPData service (http://durpdg.dur.ac.uk/, Durham UK) is an *ante litteram* data repository in the field, extracting high-level information from HEP papers or receiving direct submissions of results. The ecosystem now benefits from INSPIRE providing HEPData with full-text search of data, abstract and articles at once, while serving ancillary files with articles and issuing DOIs to this added-value material (Praczyk & Nogueras-Iso, 2012).
- The Particle Data Group (http://pdg.lbl.gov/), produces averages of thousands of measurements in the field. In synergy with INSIPRE users can now navigate from each measurement to the corresponding bibliographic record and from each record to all similar measurement, through a dedicated ontology.
- Individual HEP laboratories are also independent nodes in the ecosystem, each producing specific, institutional, information fed back to INSPIRE when relevant to a wider community. An example are results in accelerator physics as well as "experimental notes": advanced lab notebooks shared even before the preprint stage, e.g. before key conferences in the field.
- Institutional repositories as well as a multitude of websites run by research groups list theses relevant to the field. INSPIRE locates, harvests, and index them together with the other literature in the field. Theses are presented in the respective authors page of INSPIRE, aggregating the scientific output of members of the community.
- The CERN institutional repository CDS (http://cds.cern.ch/) offers collaborative authoring tool for several, thousand-strong, CERN collaborations. A two-ways institutional-repository/discipline-portal

partnership allows CDS and INSPIRE to add value for users from the different ends of the ecosystem. CDS can use integrated author-ID systems from INSPIRE, while institutional material can be immediately disseminated to the entire community.

- The wider community of conference organizers and groups recruiting new members, all leverage the role of INSPIRE as a community hub to advertise conference notices and job vacancies.

The importance of curation

A large-scale survey found that 9 in 10 HEP researchers rely uniquely on community tools such as arXiv and SPIRES/INSPIRE to access all scientific information they need (Gentil-Beccot et al. , 2009), the rest relies on Google (Scholar) which in turns indexes arXiv and INSPIRE. A key reason is INSPIRE's uniform curation process, which "normalizes" information from a variety of sources across the entire ecosystem. In particular, authors, references, researchers' affiliations and keywords are standardized and enriched, resulting in high quality metadata. The users' community relies on this specialized role of INSPIRE in the HEP repository ecosystem to assemble high-quality metadata in terms of correct reference extraction and citation counting of individual scientific artifacts. As the "user facing" part of the ecosystem, INSPIRE receives up to 50 requests per day from authors suggesting corrections of reference-related information.

Author-centric view, and the power of crowdsourcing

INSPIRE adds a dynamic, author-centric, view to the HEP ecosystem by building personal profiles which connects authors and their publications (preprints, notes, proceedings, published journal articles etc.) automatically. The profile pages include name variants, affiliation lists, frequent co-authors, keywords and subject categories, detailed citation statistics (including such metrics as the *h*-index over the entire HEP *corpus*), and author-submitted biographical data (academic genealogy, affiliation history, e-mail contacts).

Waiting for ORCID, of which we are founding members, author disambiguation is a tough nut to crack. However, users expect it to be addressed in this particular part of the ecosystem, where information is aggregated and presented, rather than the other repositories and services where information is submitted. A sophisticated author disambiguation process (Weiler et al. , 2011) has been developed that builds the foundation of a hybrid approach, based on the automated assignment and the crowd. All site users can provide feedback on author/paper connections that they believe need correction. Authors can then "claim" their papers to their profile. Researchers whose e-mail addresses are known were invited to "claim" in a pilot, and 40% did so, mostly within 24h. So far more than 3'000 researchers have participated, curating around 200'000 artifacts. Participants in the pilot then recommended this tool to their colleagues (Brooks et al., 2011). This success hinges on two factors: inclination, and motivation. First, HEP researchers are a community that has always "owned" information dissemination, and thus feels connected in making data better. Second, citation metrics in INSPIRE are used in the community for evaluation purposes, as they are felt more "complete" than those from other systems. Users therefore care of about the highest possible standard of quality.

Wider partnerships for Open Science

INSPIRE is moving beyond the HEP ecosystem into the wider Open Science environment by introducing DataCite for data management and ORCID for author identity. By exposing data from HEPData in the guise of "enhancing" publications, and assigning DOIs to these high level data sets, we have seen an emerging demand to host (and thus identify) more types of high-level data of the kind used by scientists to combine results across fields. And requests to count citations to those and integrate them in the author profile pages. By working with ORCID, arXiv, publishers and other partners on an infrastructure to interoperate the internal identifiers all parts of the ecosystem have developed, and circulate disambiguation data.

Lessons Learned

Although INSPIRE occupies a highly specialized place in a mature ecosystem of a particular discipline, we believe that the lessons we have learned in building and operating it are probably universal: co-operation across all actors, within their specific roles, and the centrality of users, allows to quickly evolve tools for sharing, and thus brings Open Science about. Being nimble, open-minded, and willing to continuously evolve and communicate are essential. Maintaining open channels of communication and engaging with practitioners is vital.

INSPIRE's experience shows that partnership across all information providers in an ecosystem and close integration, as well as collaboration with the user community yields greater innovation, which in turn spurs higher expectations from users. Our challenge is to keep up with – and accelerate – this virtuous cycle at the onset of Open Science. This is an idea that we would like to discuss with the OR2013 community, and present examples of the technical obstacles and the shared solutions for interoperability in such a complex ecosystem, as well as the strategy to find a balance between high-quality hand curation of metadata, automated approaches and the opportunities for crowdsourcing.

References:

Brooks, T. C., Carli, S., Dallmeier-Tiessen, S., Mele, S., & Weiler, H. (2011, June 15). Authormagic in INSPIRE Author Disambiguation in Scholarly Communication. Retrieved from http://journal.webscience.org/485/1/158_paper.pdf

Gentil-Beccot, A., Mele, S., Holtkamp, A., O'Connell, H. B., & Brooks, T. C. (2009). Information resources in High-Energy Physics: Surveying the present landscape and charting the future course. Journal of the American Society for Information Science and Technology, 60(1), 150–160. doi:10.1002/asi.20944

Ginsparg, P. (2011). ArXiv at 20. Nature, 476(7359), 145–7. doi:10.1038/476145a

Goldschmidt-Clermont, L. (1965). Communication patterns in high-energy physics. High Energy Physics Libraries Webzine, (6). Retrieved from http://library.web.cern.ch/library/Webzine/6/papers/1/

Heuer, R.-D., Holtkamp, A., & Mele, S. (2008). Innovation in scholarly communication: Vision and projects from High-Energy Physics. Information Services and Use, 28(2), 83–96.

O'Connell, H. B. (2000). Physicists Thriving with Paperless Publishing. Retrieved from http://arxiv.org/abs/physics/0007040

Praczyk, P., & Nogueras-Iso, J. (2012). Integrating Scholarly Publications and Research Data–Preparing for Open Science, a Case Study from High-Energy Physics with Special Emphasis on (Meta) data Semantics Research, 343, 146–157. doi:10.1007/978-3-642-35233-1_16

Weiler, H., Meyer-Wegener, K., & Mele, S. (2011). Authormagic - An Approach to Author Disambiguation in Large-Scale Digital Libraries. Proceedings of the 20th ACM Conference on Information and Knowledge Management, CIKM 2011 (pp. 2293–2297). Glasgow, UK: ACM Press.