Grant agreement no. 312788

ORCID AND DATACITE
INTEROPERABILITY NETWORK

http://odin-project.eu

D3.2Proof of Concept HEP

WP3 – Proofs of concept
V1_0 [Final]

Abstract: During its first year, the ODIN project has studied the needs and particularities of the High-Energy Physics community. Based on these needs, a set of preliminary models on data exchange and workflows has been developed.

This information will be used during the second year of the project, when cross-field requirements and parallelisms will be analysed.

SEVENTH FRAMEWORK PROGRAMME

Lead beneficiary: CERN
Date:17/10/2013
Nature: Report
Dissemination level: PU (Public)

Document Information

Grant Agreement no.	312788	**Acronym**	ODIN
Full title	ORCID and DataCite Interoperability Network		
Project URL	http://odin-project.eu		
Project Coordinator	John Kaye (BL) Address: The British Library 96 Euston Road, London NW1 2DB, United Kingdom Phone: +44 20 7412 7450 Email: john.kaye@bl.uk		

ODIN is co-funded by the EC under the e-Infrastructures Activity of the FP7 Capacities Specific Programme.

Deliverable	**Number**	3.2	**Title**	Proof of Concept HEP
Work package	**Number**	3	**Title**	Proofs of Concept

Document identifier	ODIN-WP3-DEL-0002-1_0			
Delivery date	**Contractual**	Month 11	**Actual**	
Status	Version 1_0		Final ☑ Draft ☐	
Nature	Report ☑ Prototype☐ Demonstrator ☐ Other ☐			
Dissemination Level	☑ Public ☐ Restricted to other programme participants (including the Commission Services) ☐ Restricted to a specified group (including the Commission Services) ☐ Confidential, only for consortium members (including the Commission Services)			

Authors (Partner)	Sünje Dallmeier-Tiessen, Salvatore Mele, Laura Rueda (CERN) Sergio Ruiz (DataCite) Simeon Warner (arXiv)			
Responsible Author	Laura Rueda		**Email**	laura.rueda@cern.ch
	Partner	CERN	**Phone**	+41 22 76 79092

Document Status Sheet

Issue	Date	Comment	Author
0_0	14th March '13	First draft	arXiv, CERN
0_1	27th May '13	Extended draft	arXiv, CERN, DataCite
0_2	25th June '13	Extended draft, meeting at CERN	arXiv, CERN, DataCite
0_3	22th July '13	Draft for internal review	arXiv, CERN, DataCite
0_4	26th July '13	Revised draft and proof of concept suggestions	John Kaye (British Library)
0_5	29st July '13	External CERN review	Patricia Herterich and Josh Brown (CERN)
1_0	31st July '13	Final version	arXiv, CERN, DataCite

Document Change Record

Issue	Item	Reason for Change

CONTENT

1. INTRODUCTION

The tradition of scholarly communication has existed for centuries and has not changed much in that time. Nevertheless, since the advent of the Internet, community practices are evolving and new opportunities have appeared for digital preservation, workflows and dissemination. Much more than just the text publication can be shared and open research requests that new resources are accessible.

In order to provide new services and respond to community requests, large-scale digital libraries face many challenges. Under this proof of concept, the digital library INSPIRE[1]will serve as a case study for the development of new services that facilitate sharing meaningful scholarly information. In particular, INSPIRE will address how research data and other supplementary material can be integrated and how these contents can be shared openly using permanent identifiers as a support for the interoperability of the systems.

The expression “meaningful information” refers, in particular, to the community’s needs in terms of potential use and reuse of materials. Sharing and presenting supplementary materials in INSPIRE requires the design of additional workflows in the backend and the development of new display functionalities for the frontend. How to display the information is particularly relevant in INSPIRE, as each researcher’s publication record is summarized and the HEP community heavily uses this summary for various purposes, from revision to citation tracking.

The High-Energy Physics (HEP) community is heavily committed to, and shows an increasing interest in, sharing and reusing their data. INSPIRE has already started to address these needs with an initiative to ingest research data. The first provider was HepData[2], a data repository hosted by Durham University (UK), which holds up to 7000 datasets, mostly managed manually and harvested in a non-standardized way. Further developments are necessary to standardize this communication and ease the reuse of such data [1].

The exchange between INSPIRE and HepData is an example of how research data sharing in HEP is still in its infancy. In terms of enhanced research data sharing, the near future will bring more complex data and different providers. Its treatment within the database will require special attention in terms of metadata, as information like correct and updated authorship or version tracking is crucial. The same situation replicates in many fields, where similar needs arise.

[1] http://inspirehep.net/

[2] http://durpdg.dur.ac.uk/

These forthcoming challenges will increase the intricacy of the data management and increase the requirements of the service. Moreover, the particularities of the HEP community demand the presentation of full scientific records as a whole, from preprints to published articles, now including the newest research data, with statistics on the publication output and citation metrics for each individual researcher. Those statistics are a key feature for the community, as they have a very high visibility and are used in career and research assessments. High accuracy is vital for the researchers; all materials should be citable and traceable to avoid errors.

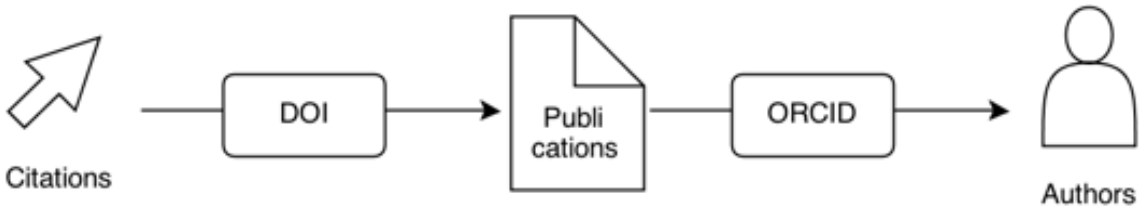


Illustration 1: Basic persistent identification schema

The use of persistent and shared identifiers between the different parts of the scholarly communication will ease the technical procedures needed to cover the community needs. Illustration 1 shows a simple schema of citation tracking, where common identifiers allow different systems to avoid tailored solutions and build reliable services. Citations based on DOIs can be extracted in any system, checked with a service provider such as DataCite[3] or CrossRef[4], while ORCID[5] iDs of the authors will appear in the metadata, and final attribution can be offered to the author.

INSPIRE serves as a case study and example to help compare fields. The complexity it faces, as an information hub for the whole HEP community, can be compared with similar challenges in different disciplines.

Most of the models and workflows developed to make INSPIRE's service more complete, reliable and extendable can serve as a base for other systems.

[3] http://datacite.org/

[4] http://crossref.org/

[5] http://orcid.org/

2. PARTICULARITIES AND GOALS

The main objective of the set of proofs of concept in the ODIN project is to identify the particularities and constraints of different fields, in order to compare them and generate common workflows for meaningful data exchange.

INSPIRE, as an aggregator of scholarly communication in the HEP field, is a hub of many resources. It has to deal with both integrating a diverse environment and offering a uniform interface for them.

2.1. The INSPIRE system and the new models

INSPIRE comprises over a million (metadata) records, half of them with Open Access full-text. It is the successor to the SPIRES bibliographic database, which served the HEP community for more than two decades. It represents a natural evolution of scholarly communication, built on successful community-based information systems, and offers a vision for information management in other fields of science.

INSPIRE aggregates different information sources, and presents a centralised web interface with value added services: search, author profiles, paper claiming, citation tracking, statistics, etc. Those services are built on top of the digital library software Invenio[6], developed at CERN. Invenio is, by design, an extendable platform where new functionalities can be developed.

The HEP community is particularly committed and participates by actively using the crowdsourced curation possibilities offered by INSPIRE. In particular, they have the opportunity to manually agree that a person wrote a paper. This process is called paper claiming and it has helped a lot to improve the automatic recognition processes.

[6] http://invenio-software.org/

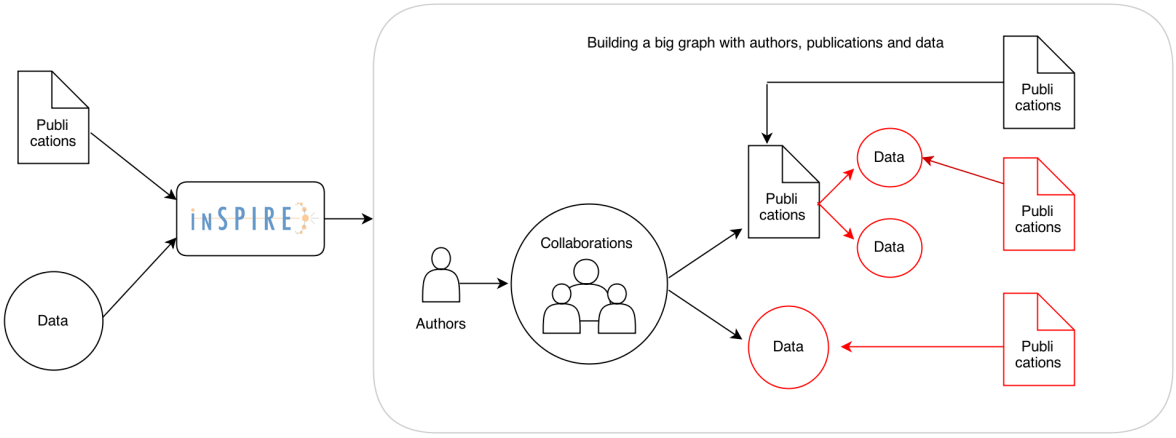


Illustration2: New models for the research output

As a proof of concept for the ODIN project, INSPIRE will be expanded to support new research and scholarly communication needs. In particular, the new models include both publications and data as research output. Illustration 2 shows in red the relationships created by these models. Data should start to be treated as a first class citizen, of equal value to text publications. This means that all the related characteristics, such as authorship, references, tracking, claiming or exchange need to be updated.

This situation becomes an excellent opportunity to redesign import mechanisms and take advantage of recently developed technologies. Many of the data included in INSPIRE are harvested using ad-hoc solutions, where widely used standards could be used. It is common that each system uses its own internal identification schema and this, sometimes, does not guarantee persistency. New developments should always be based on common identification systems and avoid non-standard solutions.

2.2. Authorship

Given the particular nature of the research in High-Energy Physics, large collaborations work together in order to design, assemble, run the particle accelerators and study the results obtained from the research process. Each person involved in the process should receive credit for his or her work.

The biggest collaborations in the field are the ATLAS[7] and CMS[8] experiments. Around three thousand researchers sign their publications, as it can be seen in Illustration 3. It shows an article published by the ATLAS collaboration.

[7] http://atlas.ch/

This amount of names is difficult to manage, not only for manual curation but also technically, as many systems are not prepared for the length of the metadata. This phenomenon is usually referred as Hyperauthorship.

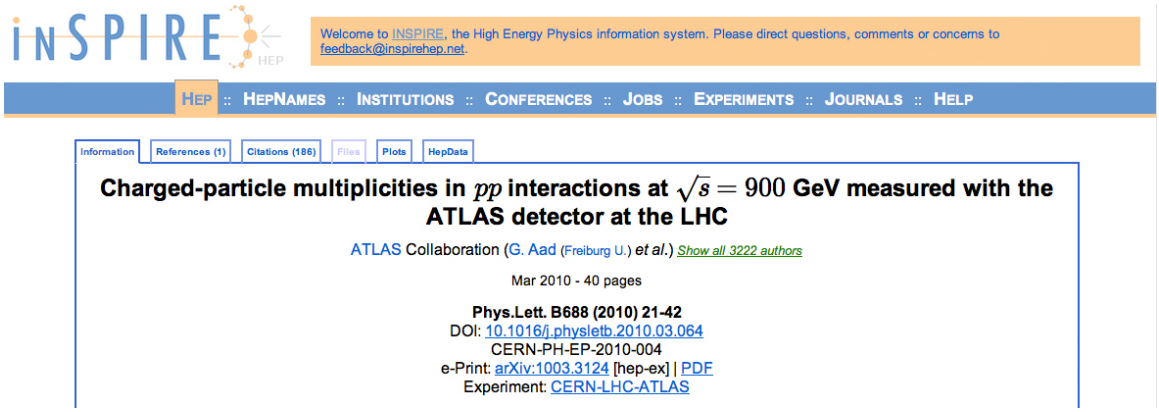
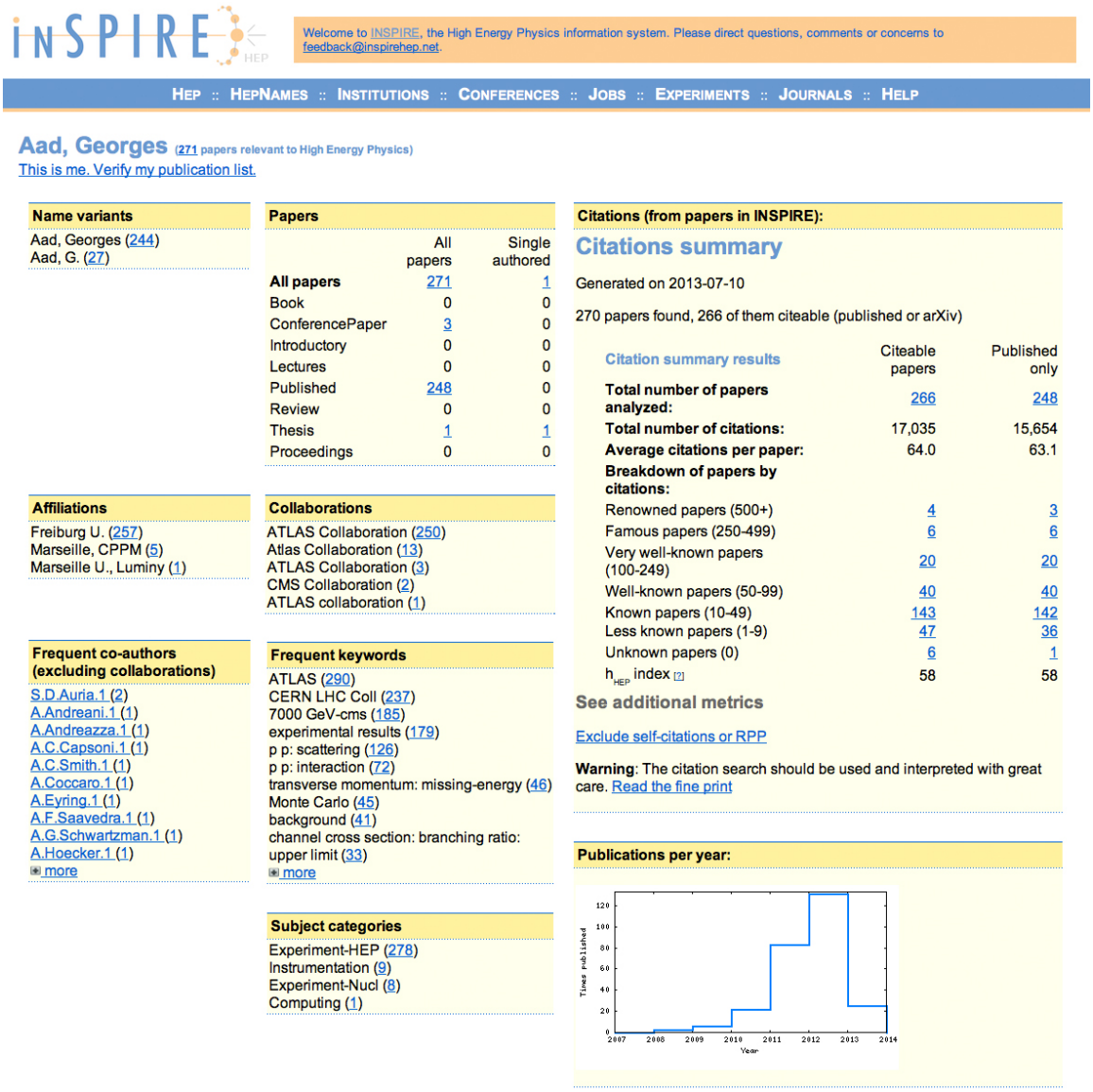


Illustration 3: ATLAS Collaboration paper with 3222 authors

[8] http://cms.web.cern.ch/



INSPIRE HEP

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

Aad, Georges (271 papers relevant to High Energy Physics)

This is me. Verify my publication list.

Name variants

Aad, Georges (244)
Aad, G. (27)

Papers

	All papers	Single authored
All papers	271	1
Book	0	0
ConferencePaper	3	0
Introductory	0	0
Lectures	0	0
Published	248	0
Review	0	0
Thesis	1	1
Proceedings	0	0

Citations (from papers in INSPIRE):

Citations summary

Generated on 2013-07-10

270 papers found, 266 of them citeable (published or arXiv)

Citation summary results	Citeable papers	Published only
Total number of papers analyzed:	266	248
Total number of citations:	17,035	15,654
Average citations per paper:	64.0	63.1
Breakdown of papers by citations:		
Renowned papers (500+)	4	3
Famous papers (250-499)	6	6
Very well-known papers (100-249)	20	20
Well-known papers (50-99)	40	40
Known papers (10-49)	143	142
Less known papers (1-9)	47	36
Unknown papers (0)	6	1
h_{HEP} index [?]	58	58

See additional metrics

Exclude self-citations or RPP

Warning: The citation search should be used and interpreted with great care. Read the fine print

Affiliations

Freiburg U. (257)
Marseille, CPPM (5)
Marseille U., Luminy (1)

Collaborations

ATLAS Collaboration (250)
Atlas Collaboration (13)
ATLAS Collaboration (3)
CMS Collaboration (2)
ATLAS collaboration (1)

Frequent co-authors (excluding collaborations)

S.D.Auria.1 (2)
A.Andreani.1 (1)
A.Andreazza.1 (1)
A.C.Capsoni.1 (1)
A.C.Smith.1 (1)
A.Coccaro.1 (1)
A.Eyring.1 (1)
A.F.Saavedra.1 (1)
A.G.Schwartzman.1 (1)
A.Hoecker.1 (1)
more

Frequent keywords

ATLAS (290)
CERN LHC Coll (237)
7000 GeV-cms (185)
experimental results (179)
p p: scattering (126)
p p: interaction (72)
transverse momentum: missing-energy (46)
Monte Carlo (45)
background (41)
channel cross section: branching ratio: upper limit (33)
more

Publications per year:

Subject categories

Experiment-HEP (278)
Instrumentation (9)
Experiment-Nucl (8)
Computing (1)

Illustration 4: Author page in INSPIRE[9]

It is necessary to keep track of all those single authors and guarantee effective linking with the references to the publications. This information among others like co-authors, statistics or keywords is presented in INSPIRE in the so-called author pages.

Illustration 4 shows an author page in INSPIRE, where information on the number of publications and citations, as well as affiliations over time, frequent co-authors,

[9] http://inspirehep.net/author/G.Aad.1/

keywords or statistics are shown. The accuracy of the data presented in the author pages is crucial for the researchers, given its high visibility.

These author profiles are generated automatically by a clustering algorithm and, in some situations, the lack of correct metadata necessitates additional manual curation.

Such issues appear because of the difficult disambiguation process. The name of the author is not enough to track him or her correctly. Many people can share the same name, different variants may be used in different publications, or it can simply change through the years (after a marriage, for example). The disambiguation algorithm takes into account affiliations, co-authors, keywords, years and other data, but they are never enough to guarantee absolute accuracy: authors move between institutions, change research groups, topics, etc.

Remarkable examples of this problem are Chinese family names. Table 1 shows a recount of the most repeated Chinese names in INSPIRE's database, with over a thousand matches.

Instances	Name
1484	Zhang, J.
1336	Li, J.
1313	Zhang, L.
1253	Liu, Y.
1246	Wu, J.
1174	Chen, A.
1135	Zhang, Z. P.
1096	Wu, S. L.
1072	Xie, Y.
1055	Wu, X.
1047	Wang, P.
1036	Shen, B. C.
1031	Su, D.

Table 1: Chinese names in INSPIRE

Moreover, all this process happen at the internal level of INSPIRE. Every time a publication is included in the system, authors are only identified using external ID schemas (such as internal ID from each service) and INSPIRE author pages need to be recalculated with the new information received.

An independent and global ID that works across disciplines, given that many HEP researchers work in adjacent disciplines in parallel, will simplify the disambiguation process.

Shared identifiers will allow different systems to exchange only the id of the author, avoiding ambiguity and errors. ORCID proposes a solution in this direction, a persistent identifier managed directly by each author, which can be shared and exchanged by external systems to build value-added services [2].

2.3. Multiple datacentres

INSPIRE acts as an aggregator of multiple sources, and builds services on top of the information received (see illustration 5). This means that it has to deal with the management of the information harvested: match duplicates, link related information (as data attached to papers) or update changes, among others.

As it is not feasible to match files, this work is done at the metadata level, comparing titles, authors and dates. Nevertheless, the completeness of the metadata and its quality remains as a big barrier.

The increasingly extensive use of Digital Object Identifiers (DOIs) and other systems of persistent identification has eased the process. Most of the systems retain the original ID of the object and share it within the metadata. By simple comparison, it is easy to spot coincidences and versions of the same content.

This kind of identification is also very helpful in order to track references. When formatted correctly, e.g. using a DOI, it becomes very simple to follow them and credit the original work.

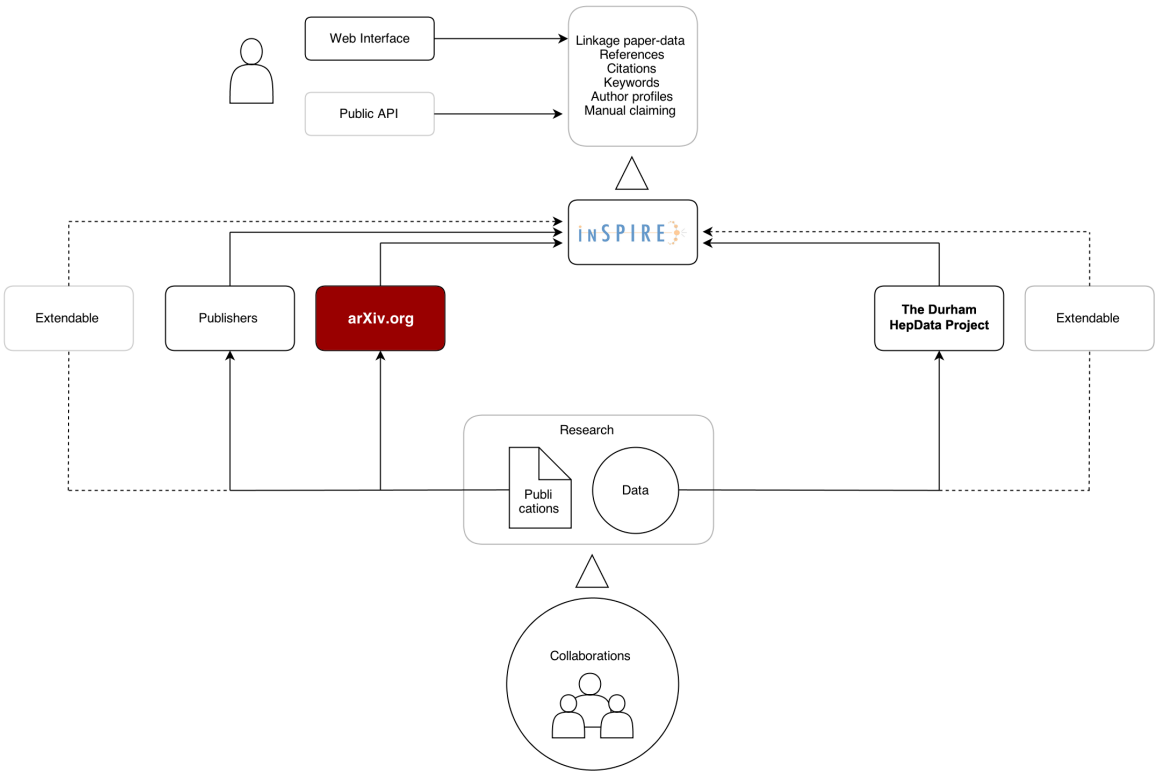


Illustration 5: INSPIRE as a content aggregator

2.4. Versioning

INSPIRE aggregates many types of material, from preprints to published articles. It generates a wide distribution of content, where INSPIRE collects and composes records for each work. Even so, the concept of different versions of the same object is currently not managed. Each record corresponds to a work as a general concept, and contains links to different services containing it.

In this way, when one of the resources change in some way, INSPIRE updates the metadata and the link to the latest version, but does not store previous versions of it.

Each user should refer to the original source of the material to obtain previous versions of the content, if available. Nevertheless, using a DOI's metadata it is possible to track versions and offer more information about them. Unfortunately, most of the material has non-extensive or non-updated metadata.

2.5. Current citation practices

Unlike other disciplines, such as Medicine or Humanities and Social Sciences, there is no tradition on primary data citation in HEP. The references are based always on previous articles, and data reuse has to be extracted via context analysis.

Previous studies, like the one carried by the GESIS – Leibniz Institute [3], have shown the enormous technical demands encountered during such reference extraction from full-text articles, and the difficulties to ensure high accuracy.

In this context, proper citation practices should be encouraged in order to simplify the extraction techniques. In the framework of this project, persistent identifiers are used as a common language, and are easy to detect, extract and process to provide reliable statistics.

Given that the HEP field has just started to introduce this concept, it should be easier to encourage and help the community to adopt standard practices. Correct data citation will allow the scientific community to reinforce the reuse of data and will simplify its dissemination and access.

3. CURRENT STATUS AND OPPORTUNITIES

The approach taken to manage both authors and data in the different systems is often very divergent. It is not possible to rebuild long established services, given technical or budget difficulties, but it is entirely necessary to offer a layer to ensure interoperability between systems.

In order to design interoperability workflows, particularly if those aim to be generic and applicable in multiple situations, it is essential to study in depth different cases and particular needs, that can be replicated among other circumstances.

3.1. Systems involved and current status

In order to generate a proof of concept that is able to tackle the different aspects and particularities of the HEP field, a set of preliminary models have been developed. They cover the whole process followed by the research output, from its generation, through its different publication steps to its final impact tracking.

The main partners for this proof of concept will be INSPIRE, as the content aggregator, arXiv, as a source for publications and ancillary files and HepData, as a source of datasets related to papers.

arXiv[10] is an archive for electronic preprints (or e-prints) operated by Cornell University. It hosts more than 800,000 scientific papers in different fields, including HEP, where almost all scientific publications are archived there.

Illustration 6 a graphical overview of the increase of publications included in arXiv since the beginning of the ‘90s.

[10] http://arxiv.org/

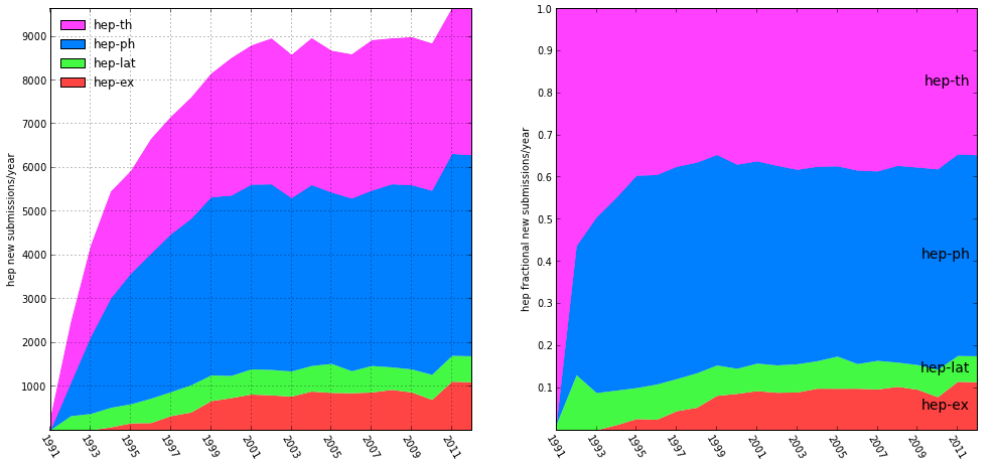


Illustration 6: High-Energy Physics publications in arXiv[11]

The shared goal between INSPIRE and arXiv is to provide the best information possible about authorship and related datasets while avoiding duplicate effort from our users. It should be based on the propagation of author identification, related files and versions of the articles in both directions. This information should also be made available as linked open data so that third parties can build other services on top of it.

There is significant overlap between the user communities of INSPIRE and arXiv, and between the datasets of each. There are more than 400.000 papers appearing in both systems, and more than 4.800 author identity associations know from the shared login mechanism.

The communication between INSPIRE and arXiv is currently based on the OAI-PHM[12] protocol. Every day, INSPIRE harvests the new articles published or updated in arXiv, under the HEP-related collection (Experiments, Lattice, Phenomenology...). The metadata received is filtered and processed to be converted into MARC-XML, so it can be easily added to INSPIRE. If the paper was published by a big collaboration, the author list is extracted.

Despite the fact that the information is checked every day, there is no version management in place. When an arXiv record is updated, INSPIRE receives the new information and can only check the main metadata fields to update its own copy, and redirect the link to the paper file to the latest version. No track or access to the previous versions is offered.

[11] http://arxiv.org/archive/stat

[12]http://www.openarchives.org/pmh/

The same issue happens with the ancillary files attached to papers that arXiv holds. Their existence is not processed by INSPIRE, so much information is lost during the harvesting process.

The HepData Project is hosted in the Durham University and has been compiling the Reaction Data Database for more than 25 years. It contains more than 7000 records of HEP scattering experiments, from a wide range of interactions. All these datasets are related to HEP published articles.

Illustration 7 shows the increase on the access demand during the last 4 years. It is important to underline that such increase is related with the start of the activity in the Large Hadron Collider (LHC) at CERN, where the two big experiments, ATLAS and CMS are located.

With the LHC working, most of the data generated at CERN has started to be demanded by multiple institutions and researchers, to carry external or associated studies. Such demand has been translated to a big increase of the use of the HepData database.

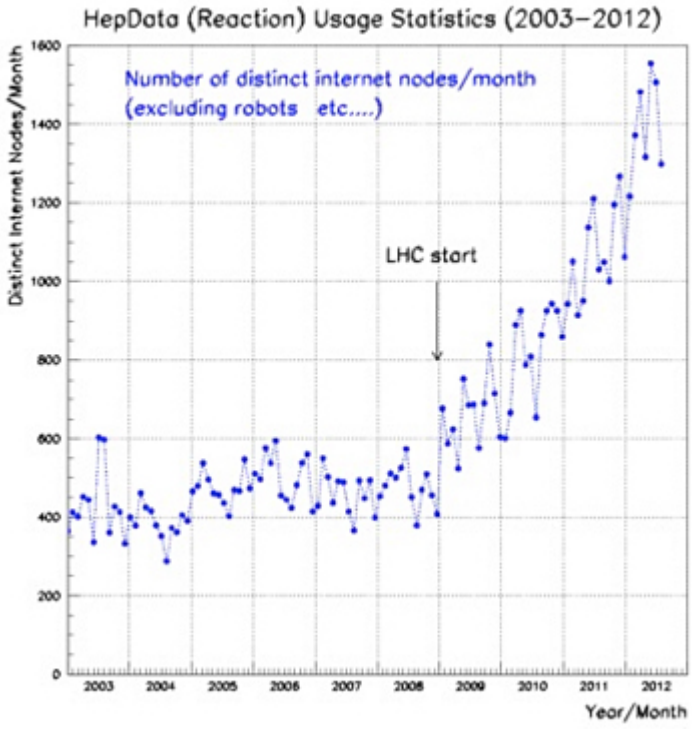


Illustration 7: HepData usage increase

INSPIRE is currently harvesting and exposing the datasets from HepData as records related to publications. INSPIRE has started as well to mint DOIs for a few of these

datasets, and will extend them to the 7000 records during the next months. DOIs will ease the citation of datasets and will enable reference tracking. It is the first step before including other data sources.

The information exchange between INSPIRE and HepData is done based on manual harvesting. INSPIRE retrieves all the records appearing in HepData and parses the HTML pages where they are shown, to extract the information and build an internal record for each dataset. The inherent characteristics of this process make it unstable: dependent on layout changes, protocol errors, heavy load, maximum number of requests, etc.

Moreover, HepData does not use any persistent identification schema or version management. During the harvesting process, INSPIRE retrieves all the datasets and matches, with the highest possible accuracy, which changes have occurred since the last update. It follows a list of records, comparing them with the previous records retrieved and tries to overcome problems, like deleted datasets or updated information.

All this setup has to be rebuilt, using a common data representation (and not HTML parsing), a set of persistent and reliable identifiers and time stamping for version tracking. Such improvements will guarantee a proper, stable and optimised data exchange.

3.2. Improvement opportunities

arXiv and INSPIRE keep a constant communication, but it is only focused on INSPIRE harvesting e-prints from arXiv. Such processes can be extended with much more information, including ancillary files, authorship or version management.

In the other direction, arXiv is not including any information from INSPIRE at the moment. Processing INSPIRE's author disambiguation results, ORCID iDs or data files can expand and improve its service. In particular, as arXiv is mostly a preprint service, the possibility to include information about the final publication of an article will be very much appreciated by the users. Such content can be generated based on INSPIRE knowledge, as easily as exchanging the DOI of the published version.

Building such interfaces for data exchange is an opportunity to offer the opportunity of profiting from the experience and databases of arXiv and INSPIRE to other systems. Third-party services should be allowed as a way of encouraging open data services and community oriented advantages.

As it will be seen in the first models, a general method for conceptual description of the information is needed, so data merging is facilitated even if the underlying schemas differ.

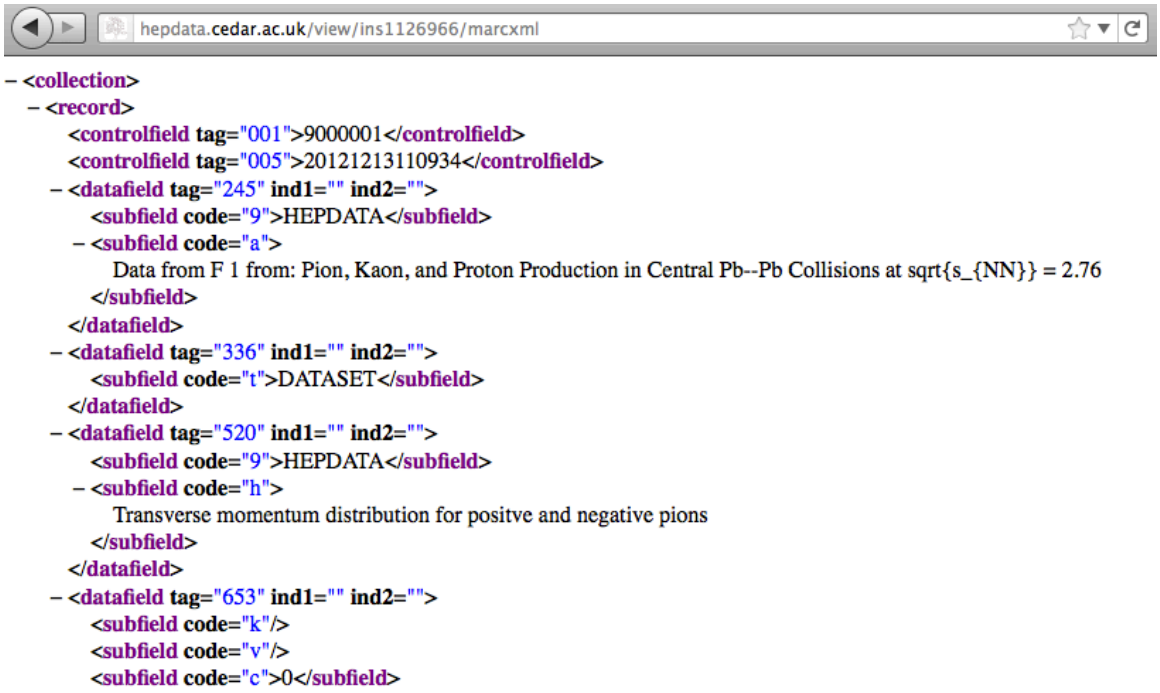


Illustration 8: MARC-XML representation for HepData

In the case of HepData, the first step has been taken. The HTML parsing system will be abandoned soon and an XML-based system will be adopted. Given the technical constraints of HepData, a simple representation has been chosen for the first tests. It is the Library of Congress' standard MARC-XML, as it is supported and easily mapped in both sides. Illustration 8 shows an example of a record in MARC-XML.

All records should be time stamped to track changes, as seen in the example. Then, once retrieved, INSPIRE can assign a DOI to each dataset and start showing and sharing the records publicly. In this way, effective reference tracking and other services are facilitated. One of the clearest benefits of these services is the possibility to develop incentive systems, based on impact rates and calculated accurately following references.

In order to implement DOIs as persistent identifiers for data, INSPIRE will use the service provided by DataCite. Given that the published articles already have a DOI assigned by the publisher, INSPIRE should focus on minting DOIs for the rest of research output, not only datasets but also plots, code snippets, etc. Such research output must be treated at the same level of the publications by establishing the same services, encouraging its reuse and tracking its impact.

This is a preliminary scenario, as the current HepData datasets harvested by INSPIRE are attached to papers. But the needs of the HEP community go in the direction of

sharing all kind of information, not only extended tables for publications but also standalone datasets. In such situation, new metadata requirements arise: In order to become reusable, the material need to be searchable in a comprehensive way, e.g. using keywords, and at the preservation level, the quality and accuracy of the metadata is indispensable.

Matching this effort to offer the technical background to facilitate the access and reuse of the data, the HEP community will be encouraged to follow standard citation practices. The preliminary designs for the data collection in INSPIRE set that all data records will have a citation suggestion, and the opportunity to export the reference in different formats, like BibTeX[13] or LaTeX[14]. Illustration 9 shows a preview of the design, where the suggestion appears as "Cite as:" with the author or collaboration, the date of publication and the DOI. Other permanent identifiers, like Handles will we shown in the same way. Then, in the bottom of the record, multiple export links will be shown.

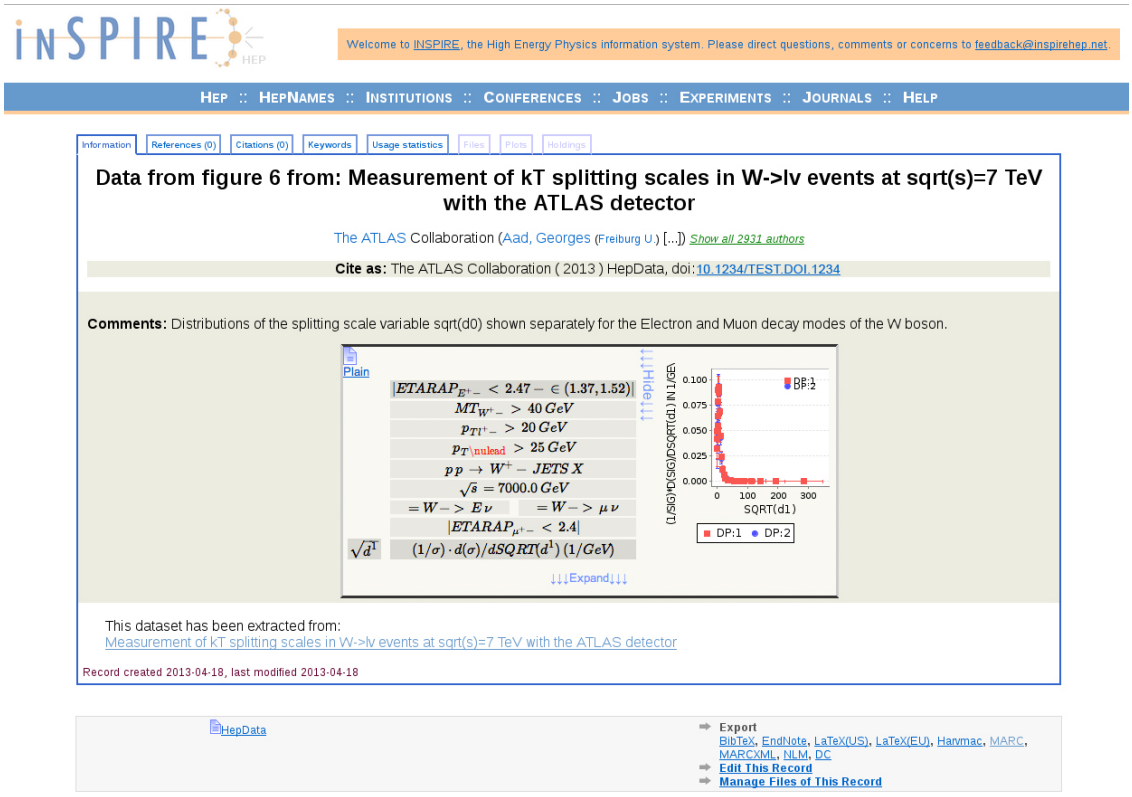


Illustration 9: Preliminary design for data records with DOI in INSPIRE

[13]http://www.bibtex.org/

[14]http://www.latex-project.org/

4. PRELIMINARY MODELS ON DATA EXCHANGE

The next design proposals have been developed during the first year of the project as an approach to generate a flexible model that could cover the general needs of the HEP community and grant successful data exchange between multiple systems. It aims also to provide a reliable interface for external systems, where added-value services can be built.

4.1. Author information exchange based on common identifiers

The main goal of this preliminary model is to guarantee that when a user associates an ORCID iD with their account in either INSPIRE or arXiv, then it should propagate through the systems without user intervention. In the same way, data associated with articles may be deposited in arXiv or may be deposited in third party systems that INSPIRE indexes. In either case the information should be surfaced in both INSPIRE and arXiv.

In order to identify authors, INSPIRE has an internal and persistent identifier called Person IDs. Those Person IDs are masked and matched to Canonical Names in a one to one relation. This is done in order to help the users, avoiding numbers and letting them use meaningful strings. It means that the users are able to access to author profiles by typing the author's name, as here: http://inspirehep.net/author/J.R.Ellis.1/.

The set Person ID/Canonical Name is built based on the signatures of the publications. A signature is the ensemble of information regarding each author of a paper, which includes its name, e-mail address and others. An automatic algorithm extracts the name of the author, his/her affiliation, co-authors, etc. and processes such information, as well as the manual claiming done to asset a publication-person relation, to generate an author page.

The Person ID/Canonical Name models the same concept as an ORCID iD: a person, the respective personal information and a list of publications. INSPIRE has already added the facility to associate ORCID iDs with Person IDS and will start collecting them in different ways. The manual option contemplates the users and administrators suggesting or connecting the ORCID iD to an author profile in INSPIRE. The automatic acquisition will collect them from the signatures of the papers and from the exchange of information with partner systems, like arXiv. A brief schema on these options can be seen in Illustration 10.

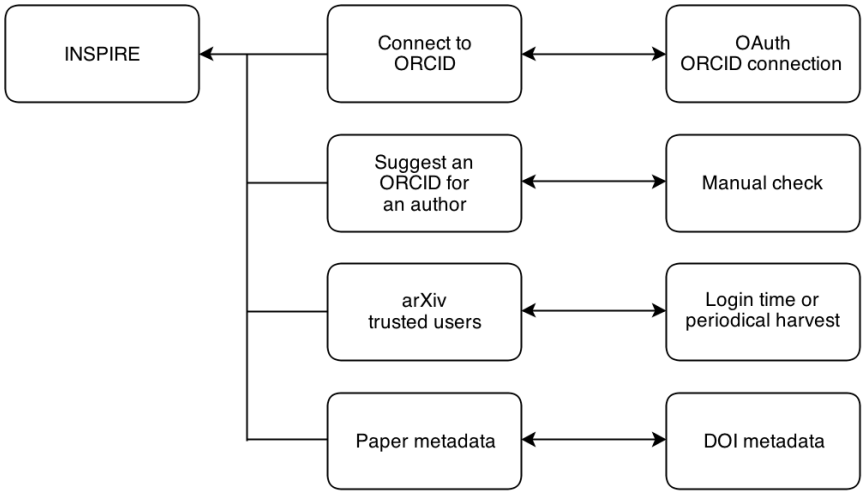


Illustration 10: ORCID acquisition in INSPIRE

It can be seen that each one of the possible ways to include ORCID iDs in INSPIRE implies a different type of interaction. The connection with ORCID is based on the open and secure protocol OAuth[15]; the iD suggestion needs a user interface capable to communicate the importance of the information; the arXiv exchange is done mostly by harvesting; and the metadata extraction is an automatic algorithm. These examples cover a wide range of interactions and can be a reference for the design of workflows in other systems.

There are almost one million papers and 7.5 million signatures on INSPIRE. Around 200.000, one fifth of the papers, have at least one manually claimed author. Those claimed papers are attached to author profiles, creating around 6.000 verified authors on INSPIRE. The implementation of ORCID iDs and the information exchange with arXiv will refine the author knowledge and the disambiguation process. A general overview of this process can be seen in Illustration 11.

[15]http://oauth.net/

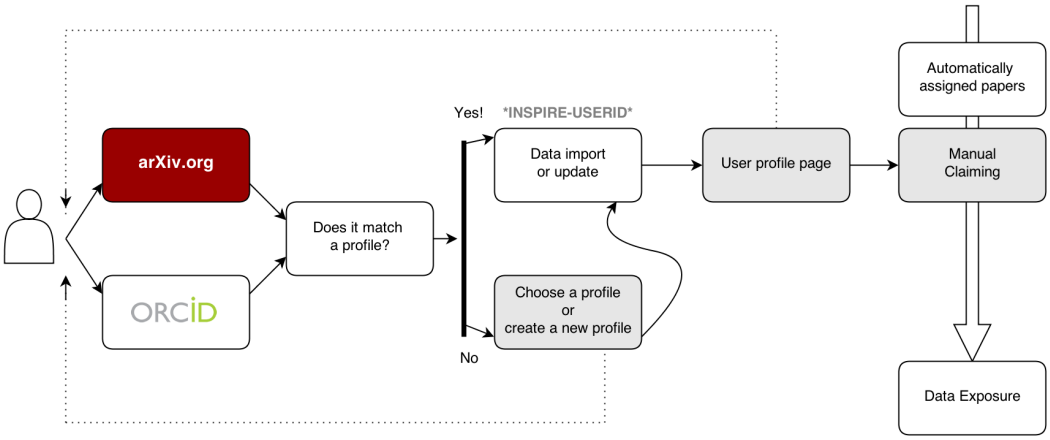


Illustration 11: Enhancement of the author information in INSPIRE

This data will be exposed in the author profile page and as part of the content shared in RDF[16] format. The next step will be an open API where a SPARQL[17] interface will be available to query this information. This will allow an easy data exchange with partners and a valuable resource to build third party value-added services.

On the arXiv side, authors are identified using URIs, such as http://arxiv.org/a/nishioka_t_1, which are created on request and tied to user accounts. arXiv will add the facility to associate an ORCID iD with a user account and this data will be exposed both on the HTML rendering of the ID page, and in data representation available from that ID (e.g.http://arxiv.org/a/nishioka_t_1.atom). As in INSPIRE, an RDF/XML data representation will be added, and it will be available via content negotiation and likely also http://arxiv.org/a/nishioka_t_1.rdf. This data can be used by INSPIRE or other third party systems to obtain ORCID iDs associated with linked accounts and enhance author recognition.

arXiv has sparse data linking authors' identifiers with papers that they have authored. There are presently about 10.000 author identifiers linking to 110.000 papers. This information will be shared via the author ID page previously mentioned and also the reverse association will be made available in an RDF/XML representation of information about each paper.

4.2. Managing data as a first class citizen

INSPIRE has started to harvest datasets from HepData and is planning to extend the same procedure to different data sources. The simplest one will be a direct data submission for the experiments at CERN, but other sources such as the Harvard

[16] http://www.w3.org/RDF/

[17] http://www.w3.org/TR/rdf-sparql-query/

Dataverse Network[18] or Figshare[19] are being considered. arXiv will be, as well, a data source. The collaboration between INSPIRE and arXiv will define workflows to be reused in the scenarios to come.

The first step to treat data at the same level as publications is to allow correct citation practices. Illustration 11 shows how the DOI assignment has been included in INSPIRE's workflows, so data is fully citable before exposing it. If the data harvested already has a DOI assigned, this intermediate step will be skipped.

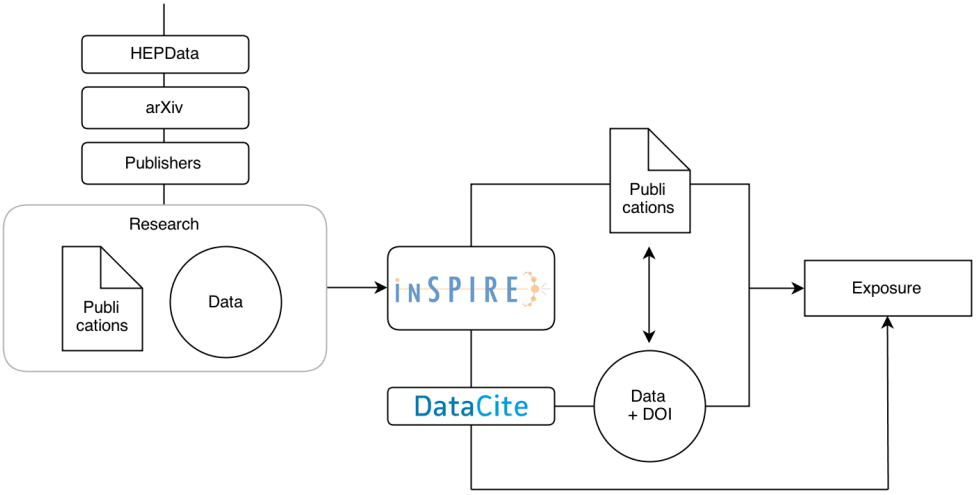


Illustration 12: DOI assignment for datasets in INSPIRE

An appropriate metadata matching has to be done, in order to provide the most extensive and useful information. In the particular case between INSPIRE and DataCite, such matching was done between the MARC21[20] standard used in INSPIRE and the DataCite Metadata schema[21].

The mandatory information to mint a DOI, also known as metadata kernel, requires a minimum set of information. After that, an extensive collection of properties helps to describe accurately the object [4].

Table 2 shows a few examples of metadata matching, concerning the kernel information and INSPIRE's information.

DataCite's Schema	INSPIRE's MARC fields

[18] http://thedata.org/

[19] http://figshare.com/

[20] http://www.loc.gov/marc/bibliographic/

[21] http://schema.datacite.org/

Identifier	0247_2, 0247_A
Creator	100__A, 700__A
Title	245__A
Publisher	260__B
Publication Year	269__C, 773__T, 961__X, 260__C

Table 2: DataCite and INSPIRE metadata match

arXiv accepts ancillary files as part of the submission process. Although it is primarily an archive and distribution service for research articles, there are limited facilities for including ancillary files of modest size (up to a few MB). Such ancillary files can include raw data for tables and plots in the article, program code, additional images, workbooks or spreadsheets. This ancillary material is only in support of research articles submitted, and no stand-alone datasets are managed.

From 2010 to 2013, arXiv collaborated on a pilot project with Data Conservancy to support remote data deposit for arXiv submissions. All data deposited as part of this pilot was moved from the Data Conservancy repository to arXiv and stored using the ancillary files mechanism, extending arXiv's data content.

As part of the ODIN project, arXiv will assign DataCite DOIs for all ancillary files and expose this information publicly, so other services can retrieve the information. In particular, INSPIRE will harvest these datasets and aggregate them as a new data source. The DOI assignment will allow INSPIRE's already in place citation tracking to follow the reference and reuse of the data initially included in arXiv, and grant credit to the authors. All this information will be shown in the author profile pages, as published data and citation statistics.

arXiv DOIs will extend INSPIRE's profiles of supported data. In this case, it will not be INSPIRE minting the DOIs for the material, they will be harvested as it is currently for papers. However, the ancillary files will be integrated in the data collection and they will be shown in the same way as the rest of data resources. This standardized interface should ensure a common environment for the users.

4.3. Ontology for Author/Paper/Data information in Inspire and arXiv

The data exchange improvements between INSPIRE and arXiv are focused on author identities (based on ORCID iDs) and data files associated with papers (especially using DataCite DOIs). This information can be expressed in RDF, as it is a clear and extensible way to exchange information. In fact, it is particularly convenient, as a SPARQL interface can be offered to query the database and build external services.

There are multiple standards that try to model scholarly communication and set a vocabulary for it. As INSPIRE records are built using the MARC21 format, a first approach was done using BIBFRAME[22]/FOAF[23], as the translation to MARC was easy and direct. In general, it is possible to match correctly many of the relations needed, but there are still few gaps.

The best alternative found as a vocabulary for the exchange is CERIF 1.6[24] (Common European Research Information Format), the standard developed by EuroCRIS. ORCID will adopt it soon, as it is the best match for their needs. INSPIRE and arXiv will explore in more depth this option in the next months, and probably move to it, as it looks it can fulfill all the needs under the same standard.

4.3.1 Classes and properties

In order to design a general model on how research output can be defined, there are three core classes to take into account:

Person/Organization/Group/Agent: The person or organization (e.g. a company or collaboration) responsible for creating any kind of work, and Instances of works.

Work: An abstract notion of a creative work, be it an article, video, dataset or code, for example. BIBFRAME defines a Work as "A resource reflecting a conceptual essence of the cataloguing resource" or FRBR as "A class whose members are an abstract notion of an artistic or intellectual creation".

Instance: A particular exemplification of the Work. In this case, BIBFRAME defines it as "A resource reflecting an individual, material embodiment of the Work".

Descriptions of Works and Instances may have different information depending on whether they are about published articles or data on its different forms. For this proof of concept purposes "papers" may be articles, books, theses, etc., and the rest of scholarly material will be "data". Data can as diverse as numerical sets, video, images, supporting documents, program-code, etc.

4.3.2 Properties of a Person

Following the idea of an RDF representation of the information, the FOAF ontology is useful to describe persons, activities and relations. Its vocabulary matches well the needs of this model.

[22] http://bibframe.org/

[23] http://www.foaf-project.org/

[24] http://www.eurocris.org/Index.php?page=CERIF-1.6&t=1

Using *foaf:Person* as the class, the identifier exposed by each system can be the URL of the author page. Identities in other systems will can be expressed as attributes and it will be up to each system to understand where it thinks there should be an *owl:sameAs* relation between INSPIRE and arXiv records that appear to be about the same person.

URI: The canonical URI of an author profile for INSPIRE and an author ID URI in arXiv

Full name and variants: *Foaf:name*, with multiple entries if there are variants of the name

Parts of name: The use of *foaf:familyName* or*foaf:givenName*can be interesting.

E-mail: Can be exposed for internal exchange of information between trusted services but never public. For example,INSPIREuses this information to let users login through arXiv with trusted sessions. *Foaf:mbox* or, in a better way,*foaf:mbox_sha1* (hashed e-mail address) cover this information.

ORCID iD: *Foaf:orcid* does not exist yet, but it will be very likely added soon. The list of relations with Works/Instances will be *foaf:made*.

This description should be extended to use *foaf:Agent* as the base class and then *foaf:Person* or *foaf:Organization/foaf:Group* for collaborations, and *foaf:Organization* for affiliations.

At internal level, arXiv is currently using heuristics to determine when some authors are not single persons but collaborations. This is done by recognizing typical strings like “Name Collaboration” or “Name Working Group”. Such information can be included as well, as it can be helpful for other systems.

4.3.3 Properties of a Work

INSPIRE’s information is mostly collected at the Work level, which is an amalgamated and/or selected description of one or more Instances. For papers, there is usually an arXiv instance, which is the latest version of the arXiv content (as commented previously, versioning is not yet managed in INSPIRE); and a publisher instance (with DOI, journal details, etc.). For data, the current instances at present come from HepData.

On the contrary, arXiv has very little notion of the abstract Work. arXiv has one or more specific instances (versions) and also holds a URI for the current “latest version” (without the version) which is what INSPIRE knows about. In some situations, arXiv may know some information about a publisher Instance. This will just cover the DOI and the bibliographic reference.

In order to exchange the different level of knowledge of each system, the idea of Work should be clearly defined. Again, *Bf:Work* class from BIBFRAME and predicates from the Dublin Core terms vocabulary can match correctly the needs of this data exchange.

Given that each work will be persistently identified by a DOI, many of the metadata can be retrieved from the DataCite service. DataCite metadata schema uses Dublin Core as its vocabulary, making easy and convenient to match fields.

The *foaf:maker* predicate can also cover the Person/Organization/Group/Agent responsible for creating the work, while *dc:creator* will be used for the authorship string.

URI: INSPIRE can provide a URI for the Work concept (e.g. http://inspirehep.net/record/849050). arXivdoes not provide any URI for this at the moment.

Title: The general title of the Work is *dct:title*. It is important to note that it may vary at the Instance level.

Creator resources: As previously explained: *foaf:maker.*

Creator string: Idem: *dct:creator.*

Abstract:*dct:abstract.*

Categories in arXiv:*arxiv:categories* as has very specific meaning to arXiv.Although other systems do not manage this characteristic, this information will be exposed by arXiv and may be useful for third parties as well as forINSPIRE in selecting items.

Publication date: Should be the publication date of the first instance known about. dct:date is defined using the standard for dates ISO8601.

Report numbers: This might be alternatively understood as different instances, but that probably is not very applicable. *dct:identifier* cannot match, in any case.

Related resources:*bf:relatedResource,*it will be more interesting to have something like "*bf:hasInstance"*as BIBFRAME has only the inverse *bf:instanceOf* to link to Instances of the Work

Related work: INSPIRE stores links from data to the associate paper.It is possible to keep the relation in both ways in RDF, e.g. *bf:relatedWork.* For INSPIRE this will be a "Work to Work" relation, but for arXiv there should be defined an"Instance to work" relation.

4.3.4 Properties of an Instance

As previously said, arXiv is focused on specific, concrete instances of a Work. Each arXiv version is an Instance. Each Instance has one or more files and may have associated data or ancillary files.

Given that the main content of arXiv is preprints, the publisher version of an article should be treated as a separate Instance, identified by the publisher DOI where available.

It is often the case that INSPIRE has both a DOI and a resolved URI to the publisher Instance. Following the DOI as the clear identifier of the Instance, it should be used as the identity and the resolved URI, if useful, can indicated via other means (could be *owl:sameAs* or using a *dct:identifier* predicate).

The management of non-persistent identifiers, as the URLs, is dangerous. This information can change or become obsolete and incorrect information will be distributed. Not only for this proof of concept but in general, it will be safer to indicate all extra URLs with dct:identifierand not with *owl:sameAs*. Nevertheless, avoiding them is recommended.

4.3.1.1. Common for all Instances

Title: Can be the same or different from the Work level: *dct:title.*

Creator resources: Should be *foaf:maker* but neither arXiv not Inspire have this information at the Instance level.

Creator string: As in Work: *dct:creator*,

Abstract: As in Work: *dct:abstract*.

Publication date: As in Work: *dct:date*(ISO8601)

Publisher: Will be dct:publisher. Note that this cannot be part of the Work definition.

License: It is important to share, *dct:license*.

Related resource: As in Work: *bf:relatedResource*.

Related work: As in Work: *bf:relatedWork*.

4.3.1.2. Instances of papers

URI: Splash page or the record or DOI for a publisher instance. For arXiv it can be considered to use the /abs page (e.g. http://arxiv.org/abs/arXiv:1003.3124)

Version: Particularly useful for arXiv, should use arxiv:version as no other vocabulary seems to match.

Other identifiers:arXiv ID, INSPIRE record ID, etc.Should use *dct:identifier* with URI if possible.

Report numbers: As for Work: *dct:identifier*.

Related work: As for Work: *bf:relatedWork*.Can be used by arXiv to indicate relation from a specific paper instance to a dataset Work.

4.3.1.3. Instances of data

URI: For this proof of concept, all data should have a DataCite DOI (used by DataCite as the resource name).

Publication date:*dct:date*. DataCite metadata schema only requires the year.

Publisher: Using the URI if possible with *dct:publisher*. For example: http://durpdg.dur.ac.uk/ for HepData.

Type: DataCite types come from *dct:type* and are defined in the schema with the next options: Collection, Dataset, Event, Film, Image, InteractiveResource, Model, PhysicalObject, Service, Software, Sound, Text.

Format: Particularly helpful for visualization management, as different treatment can be needed. Using: *dct:format*.

4.4. Simplifying the model

As seen on the previous sections, it is difficult to strictly define a common format for two different systems. Each one models different concepts in different ways, and it is necessary to use pieces of different standards to match them.

On the other hand, these problems can be bypassed in a simple way, based on the use of a common persistent identifier with a good metadata schema. In our particular example, the communication between INSPIRE and arXiv will use DOIs as identifiers for Works/Instances. Only by sharing the DOI of the object, all the metadata can be retrieved from the DataCite service, in a one and only format (based on Dublin Core in this case). Each system can match this schema with its internal system in a simple way and avoid problems.

Nevertheless, this solution is yet far from possible. The quality of the metadata stored by the original DOI creator is often much lower than the knowledge stored by each system. In the case of the data, given that is an emerging need, big efforts are being done to provide comprehensive and wide metadata. Such initiatives should ease the way of the data exchange.

Illustration 13 shows a simple schema on the information shared by each system and how external value-added services can be plugged to this interface.

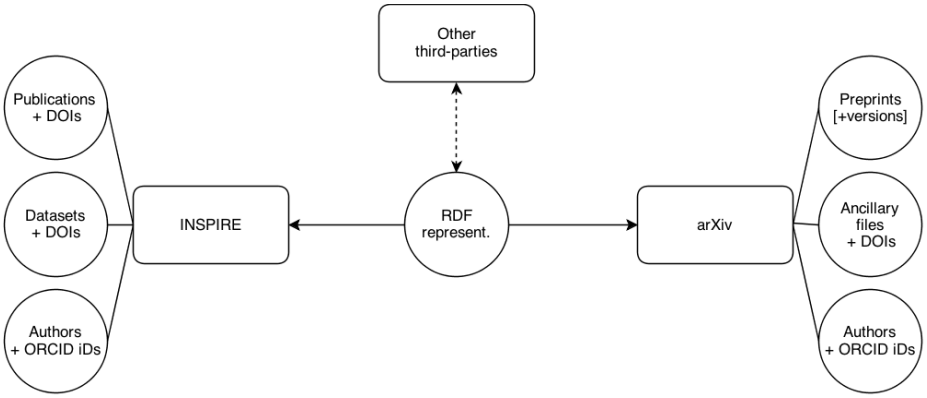


Illustration 13: General overview of the information exchange

5. CONCLUSIONS AND SECOND YEAR

During the first year, a set of needs and requests from the HEP community has been extracted, and the first steps to cover them have been established.

The collaboration between different systems has shown crucial to reach the new goals in scholarly communication. In order to do so, the development of new services should be based on data exchange, reliable standards and compatible workflows.

The use of a persistent identifier, immutable among the partners, to describe unambiguously an object or an author is the way to overcome most of the issues found. It remains to be seen which standard for data exchange is the most suitable for the particular communication between INSPIRE and arXiv.

During the second year of the project, the preliminary models and needs described in this document will be compared with the Proof of Concept carried in the Humanities and Social Sciences field. Commonalities and differences will help us to understand the general status and establish open and widely applicable workflows for simultaneous identification of contributors and datasets.

6. REFERENCES

1. Praczyk, P., Nogueras-Iso, J., Dallmeier-Tiessen, S., Whalley, M.: 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 (2012). doi:10.1007/978-3-642-35233-1_16

2. Haak, L., Fenner, M., Paglione, L., Pentz, E., Ratner, H.: ORCID: a system to uniquely identify researchers, Learned Publishing, Volume 25, Number 4, pp. 259- 264(6) (2012). doi:10.1087/20120404

3. Boland, K., Ritze, D., Eckert, K., Mathiak, B.: Identifying References to Datasets in Publications.Lecture Notes in Computer Science Volume 7489, 150-161 (2012) doi:10.1007/978-3-642-33290-6_17

4. Brase, J.: DataCite - A Global Registration Agency for Research Data, Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, pp.257-261 (2009). doi:10.1109/coinfo.2009.66