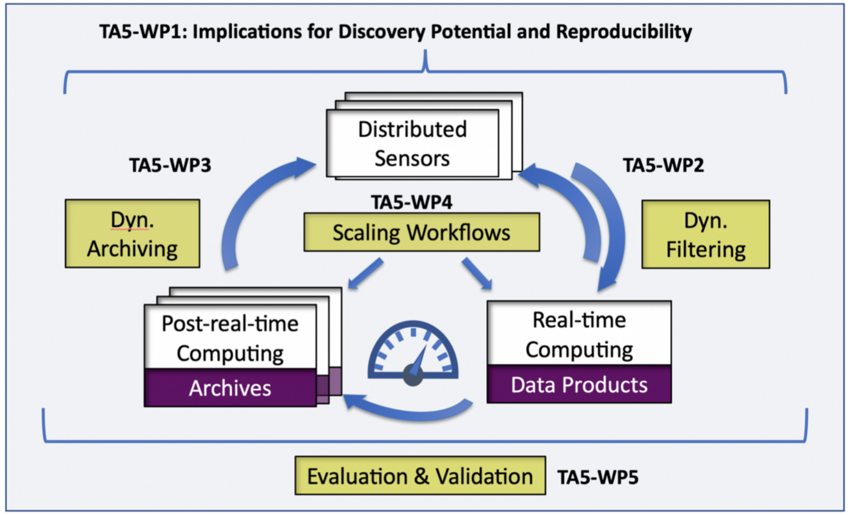
TA5 matters

Michael Kramer Andreas Redelbach

14/03/2024

TA5 - update



Dynamic Life Cycle Model

- Many activities in most work packages
- Summary of first deliverables → next slide
- Constructive **workshop on dynamic archiving** (Zeuthen, January)
    - Discussions of concepts and some technicalities
    - Future connections of „Bonn“ and „Berlin“ dynamical archives
    - Plans for discussions with experts from Virtual Observatory / HEP online calibrations
- Developments for **Machine Learning-based Pipeline for Pulsar Analysis**
    - Repository https://gitlab-p4n.aip.de/punch/ta5/wp4/ml-ppa
    - Documentation https://gitlab-p4n.aip.de/punch/ta5/wp4/ml-ppa/gitlab-profile/-/blob/main/PUNCH_interTwin_project.pdf
    - Integration tests started
- Only few activities in WP5 Evaluation and validation of instrument response & characteristics

Status of documents/deliverables

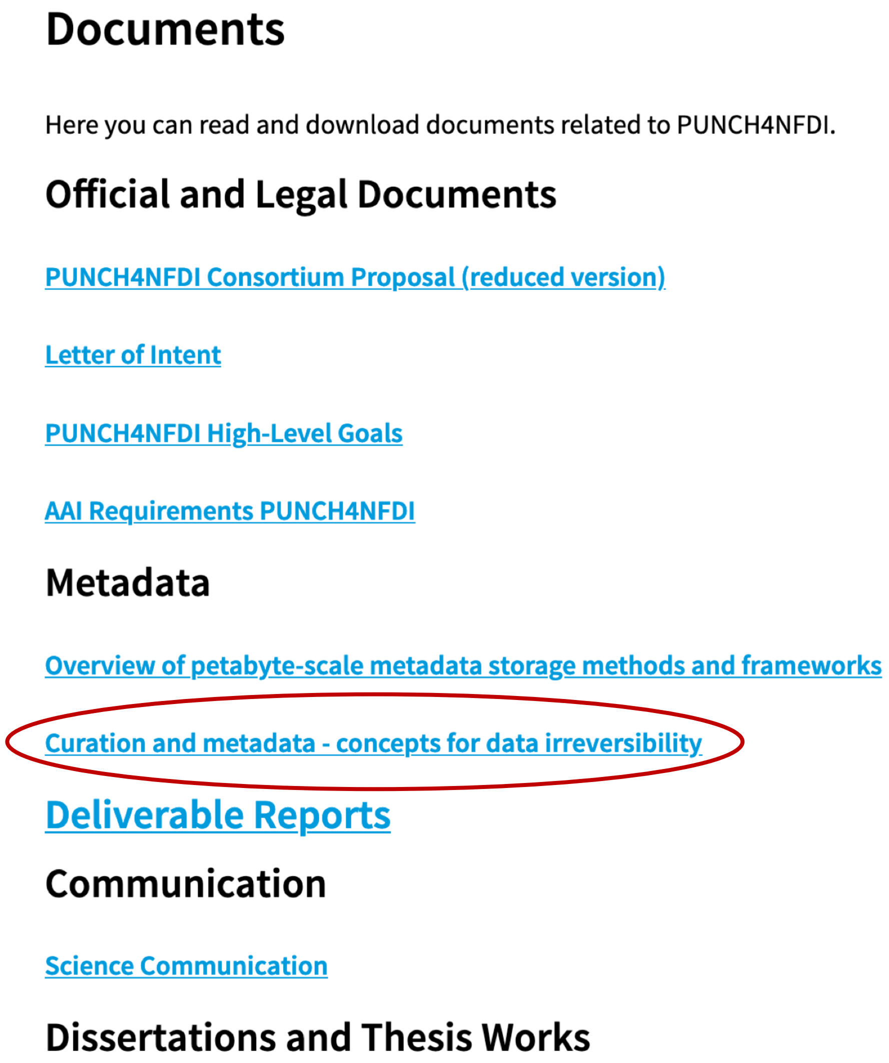
Name	Content finalized	Sent to MB/EB	Executive summary	Results page	Zenodo
D-TA5-WP2-1 Curation & metadata schemes for dynamic filtering				Added Mar 13	https://zenodo.org/records/10692169 But not listed when searching for „PUNCH4NFDI“ in Zenodo
D-TA5-WP2-2 Strategy concept for identifying highly complex (multi-parametric) signals in huge data streams					
D-TA5-WP3-1 Specifying the concept of a dynamic archive					
D-TA5-WP1-1 Report on impact of on-line filtering on discovery potential					

→ Goal: Deliverables in Zenodo soon, following PUNCH Publication Policy

Sending documents to MB
Publication via Zenodo and Results page

Pdfs in the intranet:
https://gitlab-p4n.aip.de/punch/intra-docs-content/-/tree/master/files/TA5/Documents_deliverables

Results page



Documents

Here you can read and download documents related to PUNCH4NFDI.

Official and Legal Documents

[PUNCH4NFDI Consortium Proposal (reduced version)]

[Letter of Intent]

[PUNCH4NFDI High-Level Goals]

[AAI Requirements PUNCH4NFDI]

Metadata

[Overview of petabyte-scale metadata storage methods and frameworks]

[Curation and metadata - concepts for data irreversibility]

Deliverable Reports

Communication

[Science Communication]

Dissertations and Thesis Works

Where to integrate other TA5 documents?

https://results-preview-punch4nfdi.sirrah.aip.de/?md=/docs/Documents/documents.md

Publication policy

Inform the Publication Committee about the planned publication

TA leads will notify the MB which then has one week time to review

The Publication Committee should … preferably **cover all Task Areas**.

→ Need ideally 2 persons from TA5!

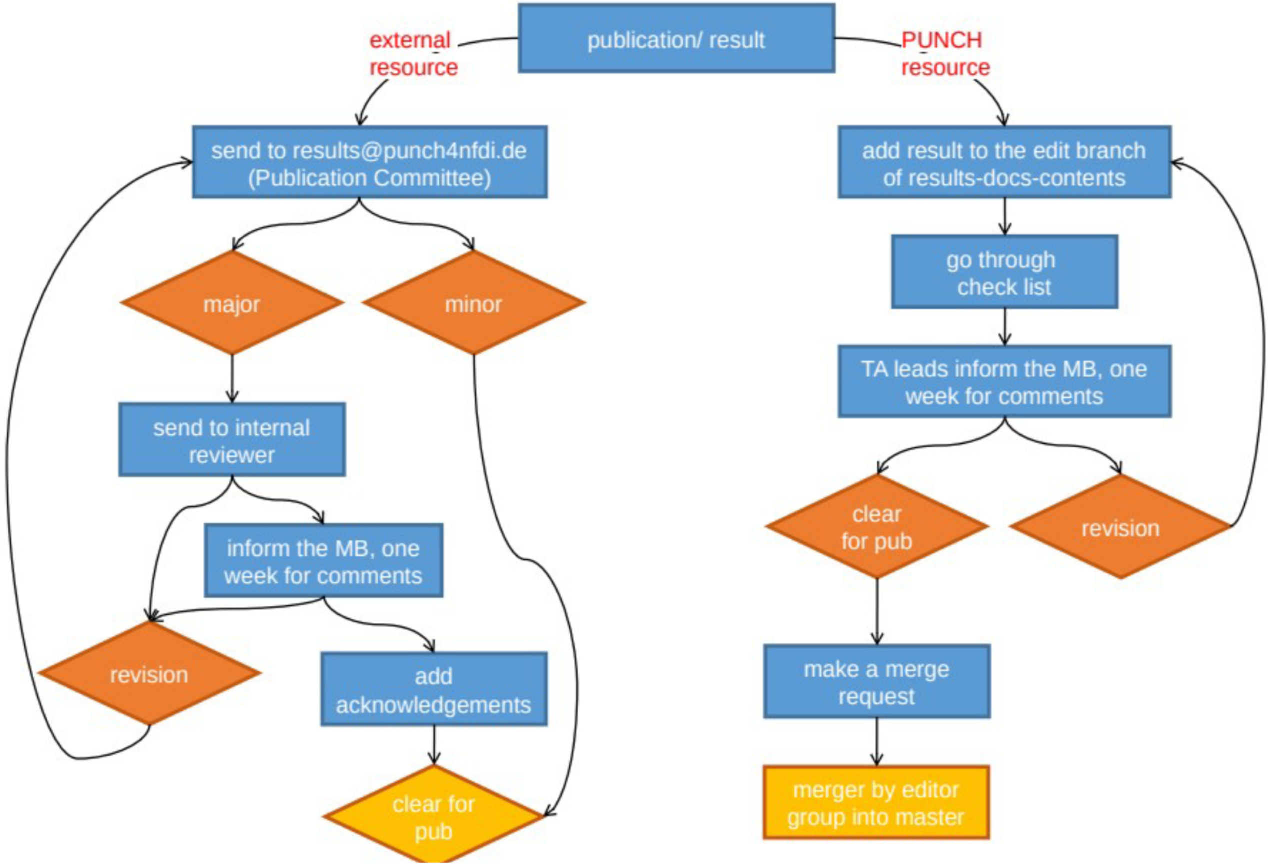

publication/ result
external resource
PUNCH resource
send to results@punch4nfdi.de (Publication Committee)
major
minor
send to internal reviewer
inform the MB, one week for comments
revision
add acknowledgements
clear for pub
add result to the edit branch of results-docs-contents
go through check list
TA leads inform the MB, one week for comments
clear for pub
revision
make a merge request
merger by editor group into master

Figure 1: Representation of the workflow for publications in PUNCH4NFDI

https://intra.punch4nfdi.de/files/PUNCH_publication_policy_v1.0.pdf

Interactions with other TAs

TA 2

- Contributions to / use cases for Compute4PUNCH or Storage4PUNCH (under discussion): Are there option to share larger data, on longer timescales (radioastronomy data sets)?

  → **first store only data of "a few TB" on the existing infrastructure** to start some tests etc.

  As soon as we need to have more storage permanently, we should also consider integrating more ressources

  Useful links: https://gitlab-p4n.aip.de/punch/intra-docs-content/-/blob/master/docs/TA2/WP1/StoragePrototyping.md
  https://gitlab-p4n.aip.de/punch/intra-docs-content/-/blob/master/docs/TA2/WP2/Compute4PUNCH_Documentation_Users.md
  https://gitlab-p4n.aip.de/compute4punch/tutorials/c4p_lofar_tutorial_general_meeting_09_2022
- ML-PPA project: Tests using containers for digital twin simulations at Jülich

TA 3

- Statistical approaches and information theory: Possible overlaps with TA3-WP1 Statistical methods – meeting to be coordinated → **need to suggest a date**
- Machine Learning projects – feedback sent

Connections to NFDI

NFDI talk:

Interest from Jakob Nordin, signed up for Oct 21

→ Confirmed for October 21

ML on FPGAs:

Discussions for follow-up of meeting (last year at DESY), possibly jointly with XFEL/DAPHNE

In person workshop and/or tutorial as option

Workshop of new **FPGA Developers' Forum at CERN on 11-13 June 2024** https://indico.cern.ch/event/1381060/

NFDI sections/services:

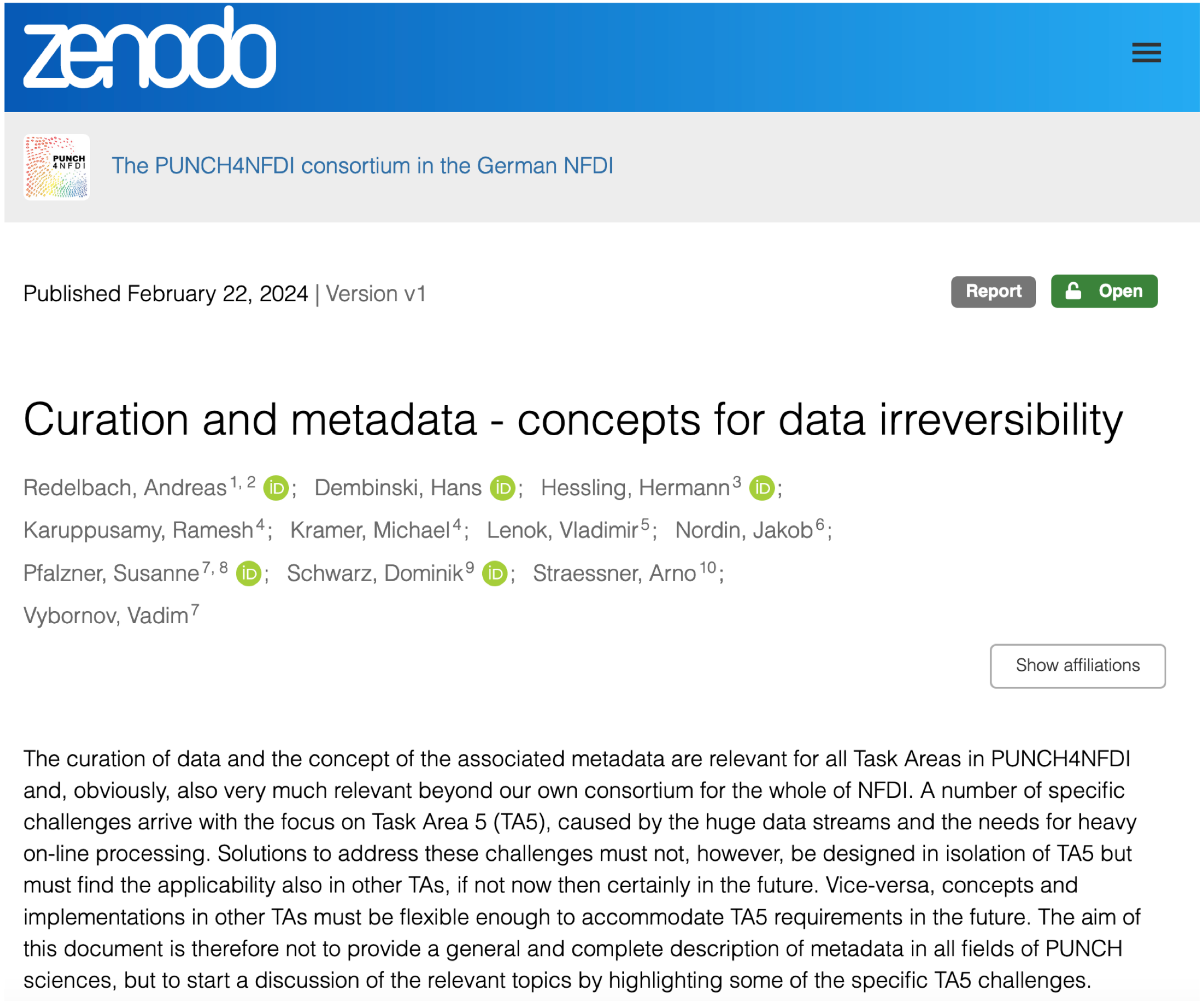
Participation to be discussed in TA5 meeting

Support for PUNCH Multi-Cloud efforts

Backup

Summary:

We have described (use) cases where limitations for existing data processing levels will be too restrictive for future extended layers or branches due to more complex workflows. Capturing the workflows of dynamic filtering/archiving shall finally enable as much reproducibility and validations as possible. Therefore, metadata must include a complete description of all algorithms involved in the pipelines/workflows.



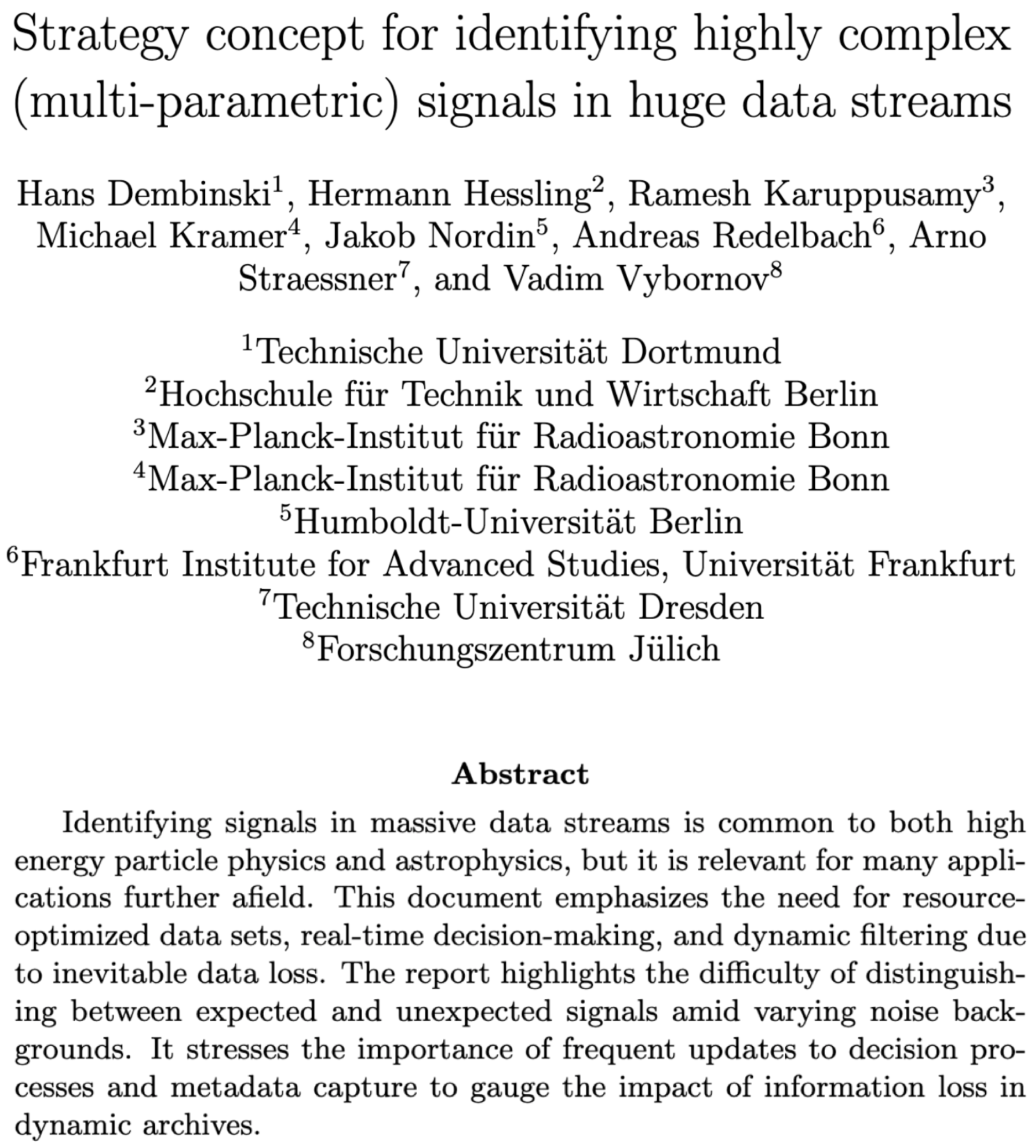
The PUNCH4NFDI consortium in the German NFDI

Published February 22, 2024 | Version v1

Curation and metadata - concepts for data irreversibility

Redelbach, Andreas[1, 2]; Dembinski, Hans; Hessling, Hermann[3]; Karuppusamy, Ramesh[4]; Kramer, Michael[4]; Lenok, Vladimir[5]; Nordin, Jakob[6]; Pfalzner, Susanne[7, 8]; Schwarz, Dominik[9]; Straessner, Arno[10]; Vybornov, Vadim[7]

The curation of data and the concept of the associated metadata are relevant for all Task Areas in PUNCH4NFDI and, obviously, also very much relevant beyond our own consortium for the whole of NFDI. A number of specific challenges arrive with the focus on Task Area 5 (TA5), caused by the huge data streams and the needs for heavy on-line processing. Solutions to address these challenges must not, however, be designed in isolation of TA5 but must find the applicability also in other TAs, if not now then certainly in the future. Vice-versa, concepts and implementations in other TAs must be flexible enough to accommodate TA5 requirements in the future. The aim of this document is therefore not to provide a general and complete description of metadata in all fields of PUNCH sciences, but to start a discussion of the relevant topics by highlighting some of the specific TA5 challenges.



Strategy concept for identifying highly complex (multi-parametric) signals in huge data streams

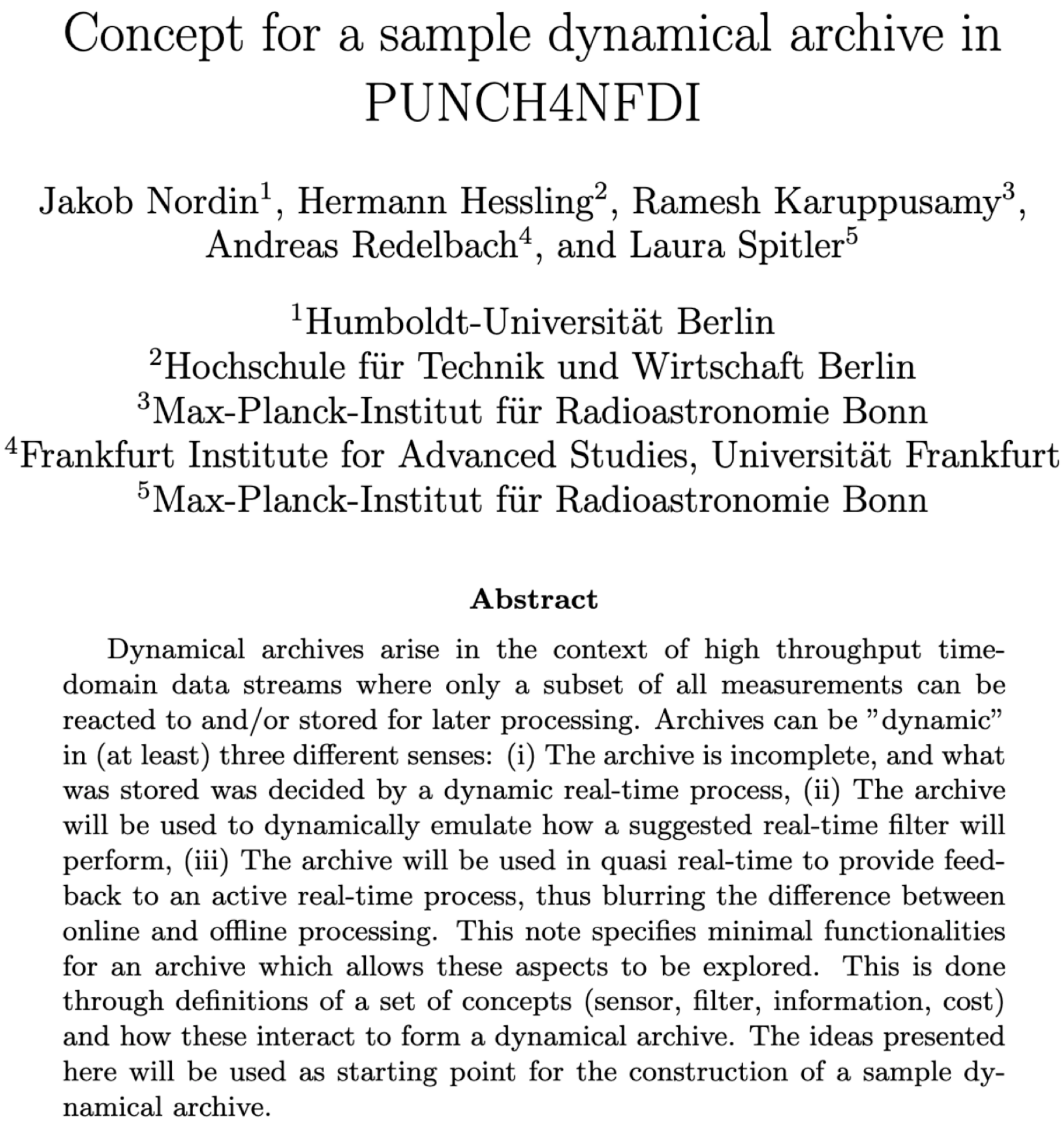
Hans Dembinski[1], Hermann Hessling[2], Ramesh Karuppusamy[3], Michael Kramer[4], Jakob Nordin[5], Andreas Redelbach[6], Arno Straessner[7], and Vadim Vybornov[8]

[1]Technische Universität Dortmund
[2]Hochschule für Technik und Wirtschaft Berlin
[3]Max-Planck-Institut für Radioastronomie Bonn
[4]Max-Planck-Institut für Radioastronomie Bonn
[5]Humboldt-Universität Berlin
[6]Frankfurt Institute for Advanced Studies, Universität Frankfurt
[7]Technische Universität Dresden
[8]Forschungszentrum Jülich

Abstract

Identifying signals in massive data streams is common to both high energy particle physics and astrophysics, but it is relevant for many applications further afield. This document emphasizes the need for resource-optimized data sets, real-time decision-making, and dynamic filtering due to inevitable data loss. The report highlights the difficulty of distinguishing between expected and unexpected signals amid varying noise backgrounds. It stresses the importance of frequent updates to decision processes and metadata capture to gauge the impact of information loss in dynamic archives.

https://gitlab-p4n.aip.de/punch/intra-docs-content/-/blob/master/files/TA5/Documents_deliverables/PUNCH4NFDI_TA5_concept_datastreams.pdf



Concept for a sample dynamical archive in PUNCH4NFDI

Jakob Nordin[1], Hermann Hessling[2], Ramesh Karuppusamy[3], Andreas Redelbach[4], and Laura Spitler[5]

[1]Humboldt-Universität Berlin
[2]Hochschule für Technik und Wirtschaft Berlin
[3]Max-Planck-Institut für Radioastronomie Bonn
[4]Frankfurt Institute for Advanced Studies, Universität Frankfurt
[5]Max-Planck-Institut für Radioastronomie Bonn

Abstract

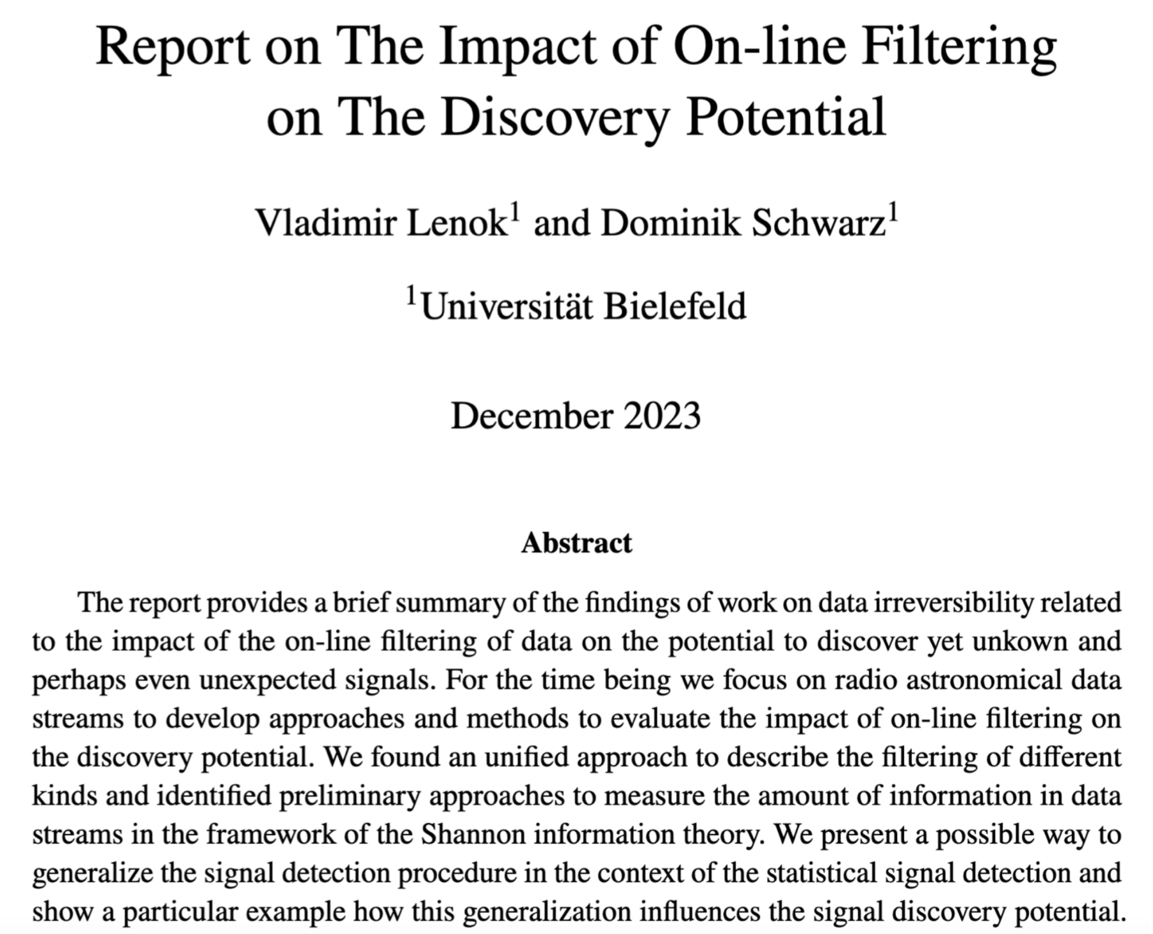
Dynamical archives arise in the context of high throughput time-domain data streams where only a subset of all measurements can be reacted to and/or stored for later processing. Archives can be ”dynamic” in (at least) three different senses: (i) The archive is incomplete, and what was stored was decided by a dynamic real-time process, (ii) The archive will be used to dynamically emulate how a suggested real-time filter will perform, (iii) The archive will be used in quasi real-time to provide feedback to an active real-time process, thus blurring the difference between online and offline processing. This note specifies minimal functionalities for an archive which allows these aspects to be explored. This is done through definitions of a set of concepts (sensor, filter, information, cost) and how these interact to form a dynamical archive. The ideas presented here will be used as starting point for the construction of a sample dynamical archive.

https://gitlab-p4n.aip.de/punch/intra-docs-content/-/blob/master/files/TA5/Documents_deliverables/PUNCH4NFDI_TA5_dynamic_archive.pdf

Executive summary:

Due to large data volumes and rates of modern physics and astrophysics experiments and observatories , the data must undergo some filtration procedures, so that it is possible to store them for further analysis. Unavoidably, such a filtration leads to loss of information originally obtained by the instrument.

The document provides a theoretical basis for quantifying the terms “amount of information” and the “information loss”. Namely, the report describes how the theories of information and statistical signal processing can be used in physical and astrophysical contexts. The studies have been based on at a particular example, the data stream provided by a single radio telescope to be generalized in the future.



Report on The Impact of On-line Filtering on The Discovery Potential

Vladimir Lenok[1] and Dominik Schwarz[1]

[1]Universität Bielefeld

December 2023

Abstract

The report provides a brief summary of the findings of work on data irreversibility related to the impact of the on-line filtering of data on the potential to discover yet unkown and perhaps even unexpected signals. For the time being we focus on radio astronomical data streams to develop approaches and methods to evaluate the impact of on-line filtering on the discovery potential. We found an unified approach to describe the filtering of different kinds and identified preliminary approaches to measure the amount of information in data streams in the framework of the Shannon information theory. We present a possible way to generalize the signal detection procedure in the context of the statistical signal detection and show a particular example how this generalization influences the signal discovery potential.

https://gitlab-p4n.aip.de/punch/intra-docs-content/-/blob/master/files/TA5/Documents_deliverables/D_TA5_WP1_1-20231218.pdf