Metadata extraction from raw astroparticle data of TAIGA experiment

Elena Korosteleva, Alexandr Kryukov, Andrey Mikhailov, Minh-Duc Nguyen, Alexey Shigarov

3rd Int. Workshop on Data Life Cycle in Physics Experiments

Irkutsk, Russia

April 3-5, 2019

Outline

- Motivation
- Background
- Metadata in TAIGA raw data
- Concept of Metadata Extractor
- Conclusion

Motivation

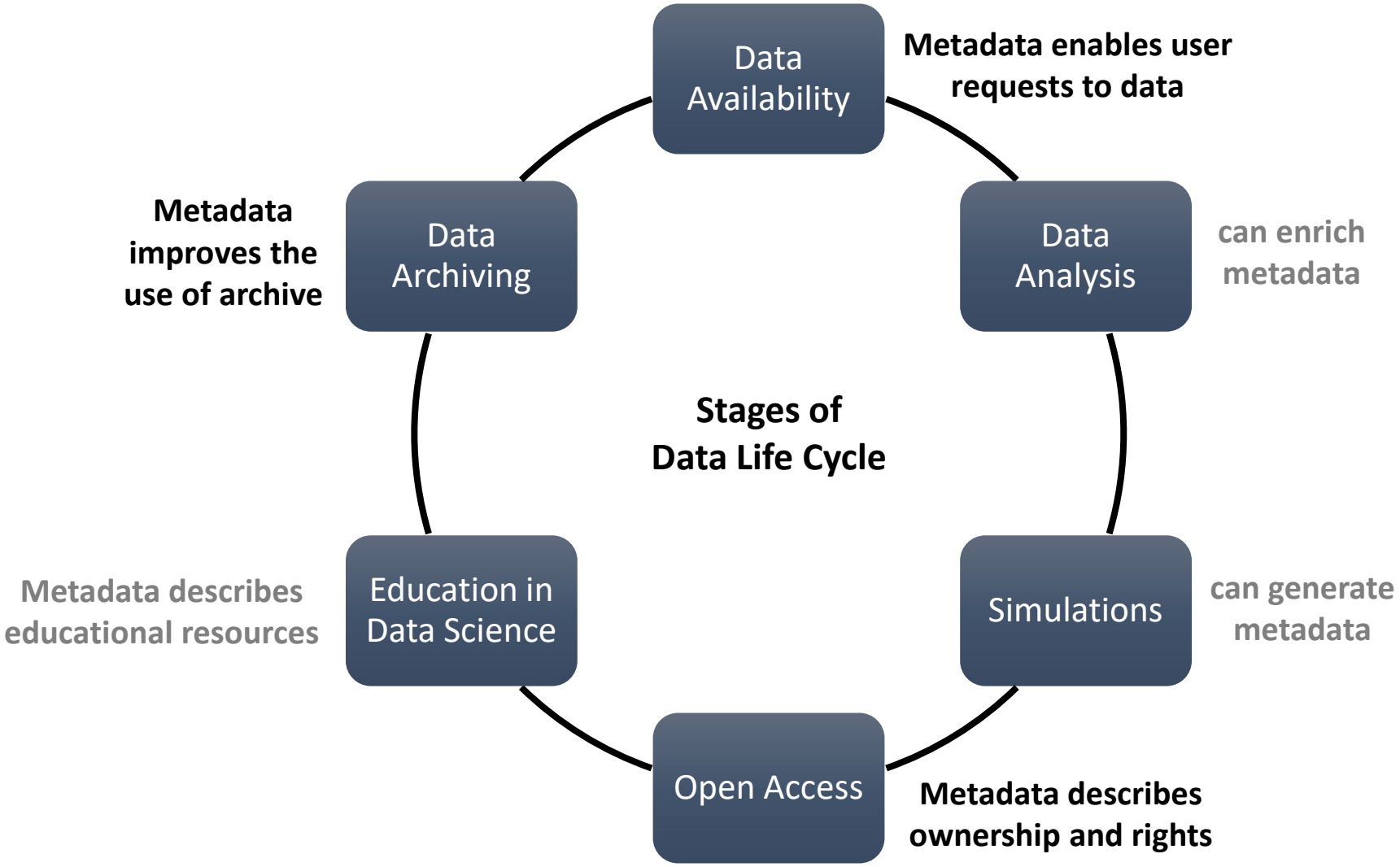
"Metadata is **data about data**"

"**Metadata** *enables and improves use of that data*"
https://guides.lib.unc.edu

Metadata

Data

Motivation


Data Availability
Metadata enables user requests to data
Data Analysis
can enrich metadata
Simulations
can generate metadata
Open Access
Metadata describes ownership and rights
Education in Data Science
Metadata describes educational resources
Data Archiving
Metadata improves the use of archive
Stages of Data Life Cycle

Motivation

- **Metadata**: *the status quo*
  - Hidden MD in file names & package headers
  - No an unified access to hidden MD
  - No an unified terminology
  - No an unified storage of MD
- **Challenges**
  - What is *metadata* for our case?
  - How to extract metadata from raw data?
  - How to store MD

Background

- **Metadata extraction**
  - characterization of ***digital objects***
  - deriving representation information about a ***digital object*** significant for purposes of classification, analysis, and use
- **Digital object**
  - “A **digital object** is composed of structured sequence of bits/bytes. … The bit sequence realizing the object can be identified & accessed by a unique and persistent identifier or by use of referencing attributes describing its properties” [1]
  - “**Digital object** is … machine-independent data structure consisting of one or more elements in digital form that can be parsed by different information systems …” [1]
  - [1] https://www.rd-alliance.org/group/data-foundation-and-terminology-wg/post/community-discussion-definition-digital-object.html

Background

- **Tools for harvesting metadata from binary files**
  - NLNZ Metadata Extraction Tool (http://meta-extractor.sourceforge.net)
  - JHOVE2 (https://bitbucket.org/jhove2/main/wiki/Home)
  - FITS (https://projects.iq.harvard.edu/fits)
  - GNU Libextractor (https://www.gnu.org/software/libextractor)
- **Functionality**
  - Support some wide-spread file formats (e.g. JPEG, MP3, ZIP)
  - Can be extended by plug-ins or modules for processing specific formats
  - Store extracted metadata in XML, JSON, or delimited text files

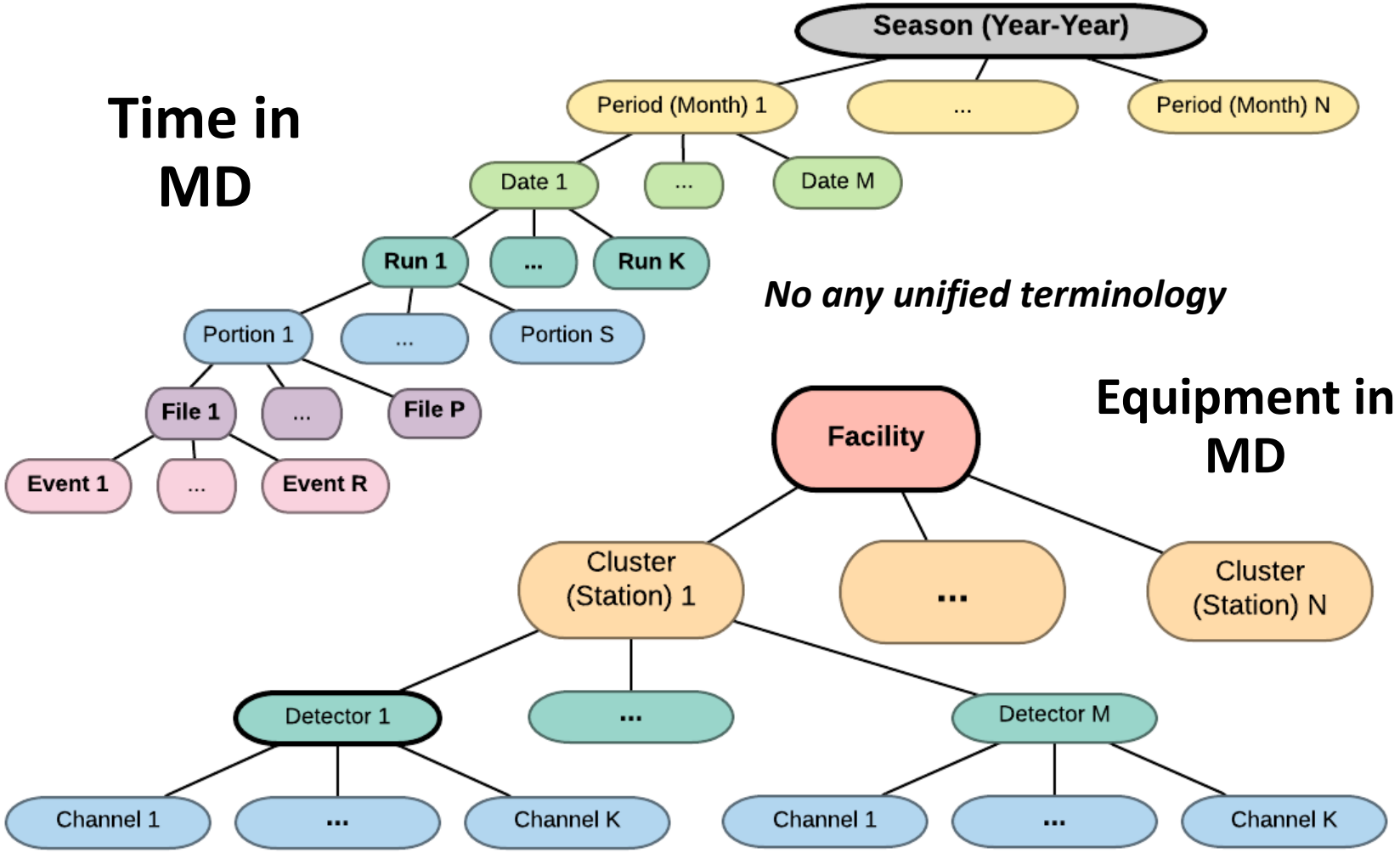
Background

- Workflow for the **characterization of digital objects** (*JHOVE2's* point of view)

Identification	Validation	Extraction	Assessment
determining the presumptive format of a digital object	determining the level of conformance to the object's format	deriving metadata of a digital object significant for purposes of classification, analysis, and use	determining the level of acceptability of a digital object for a specific use on the basis of locally-defined policies

What Metadata We Can Extract from TAIGA Raw Data

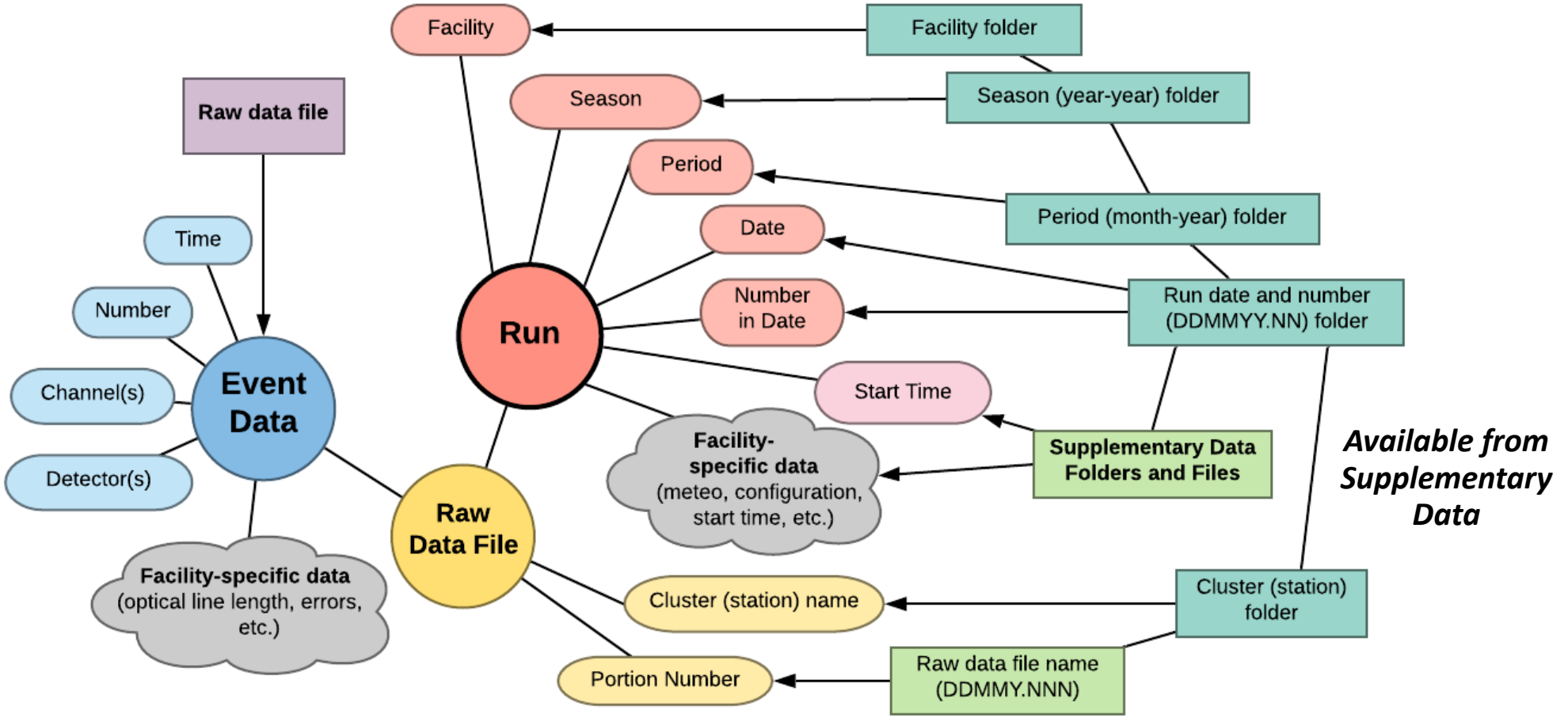
Metadata hidden in TAIGA Raw Data



Metadata hidden in TAIGA Raw Data

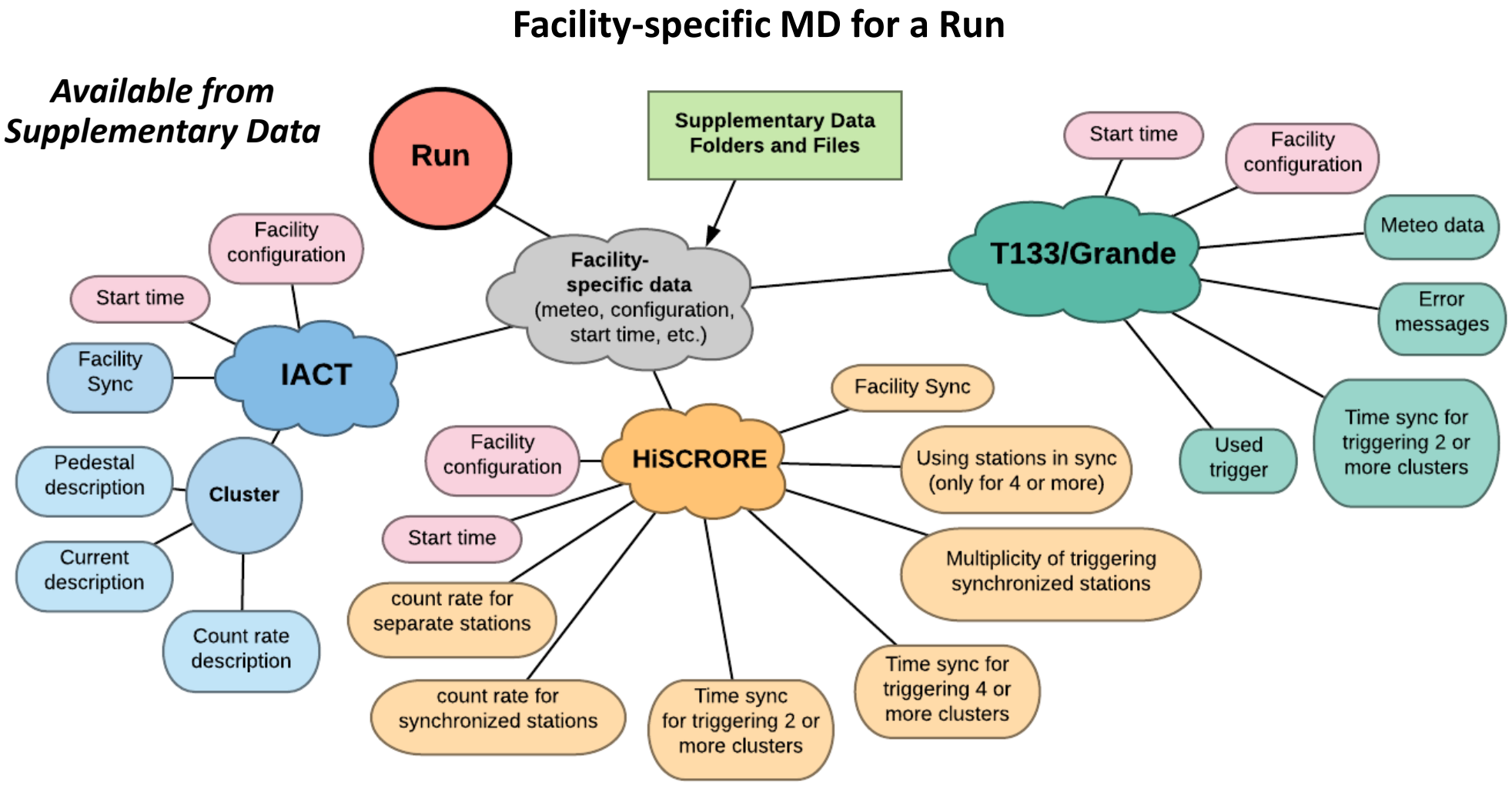
Available From Raw Data Files

Available from Folder & File Names



Available from Supplementary Data

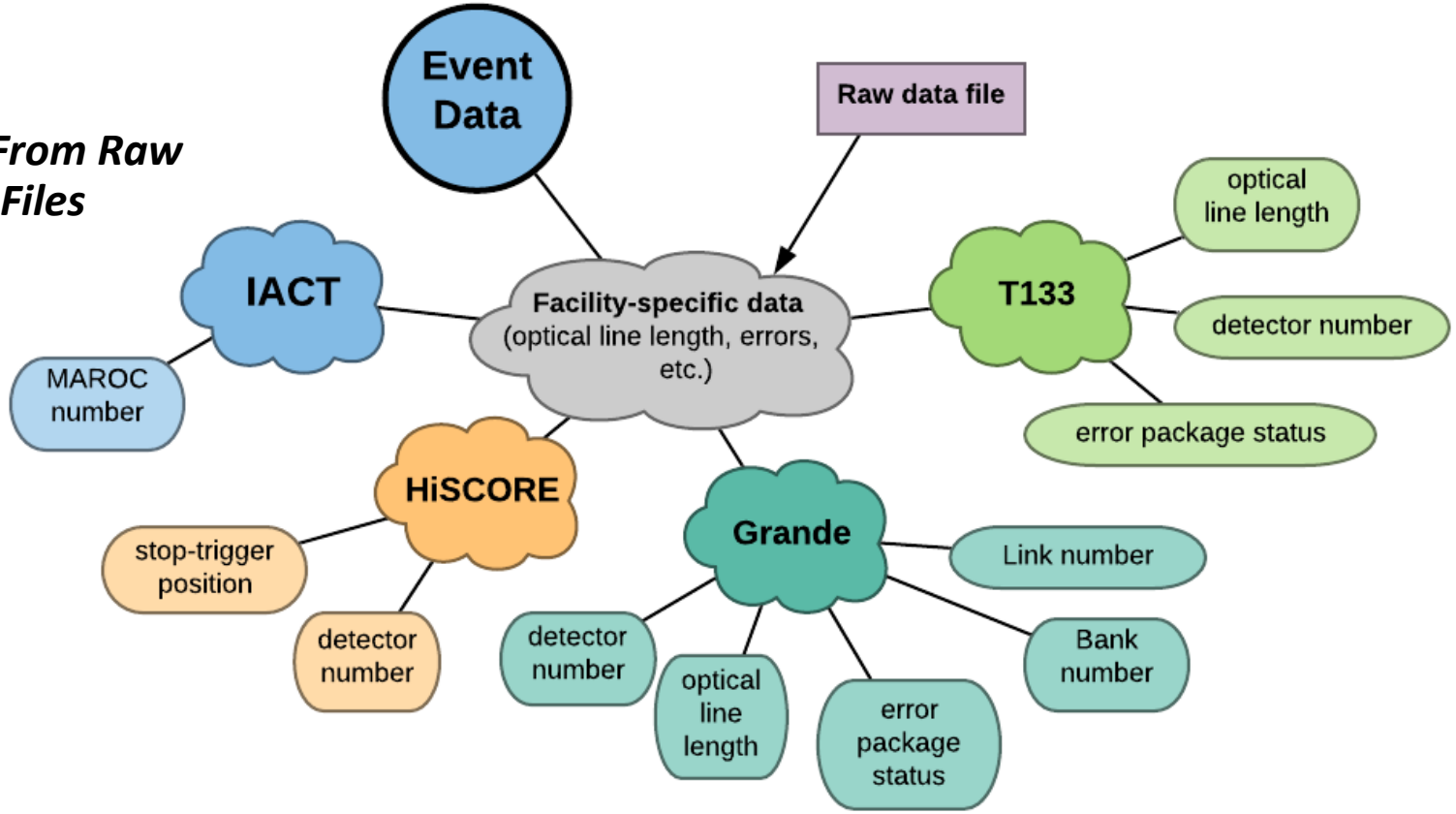
Metadata hidden in TAIGA Raw Data



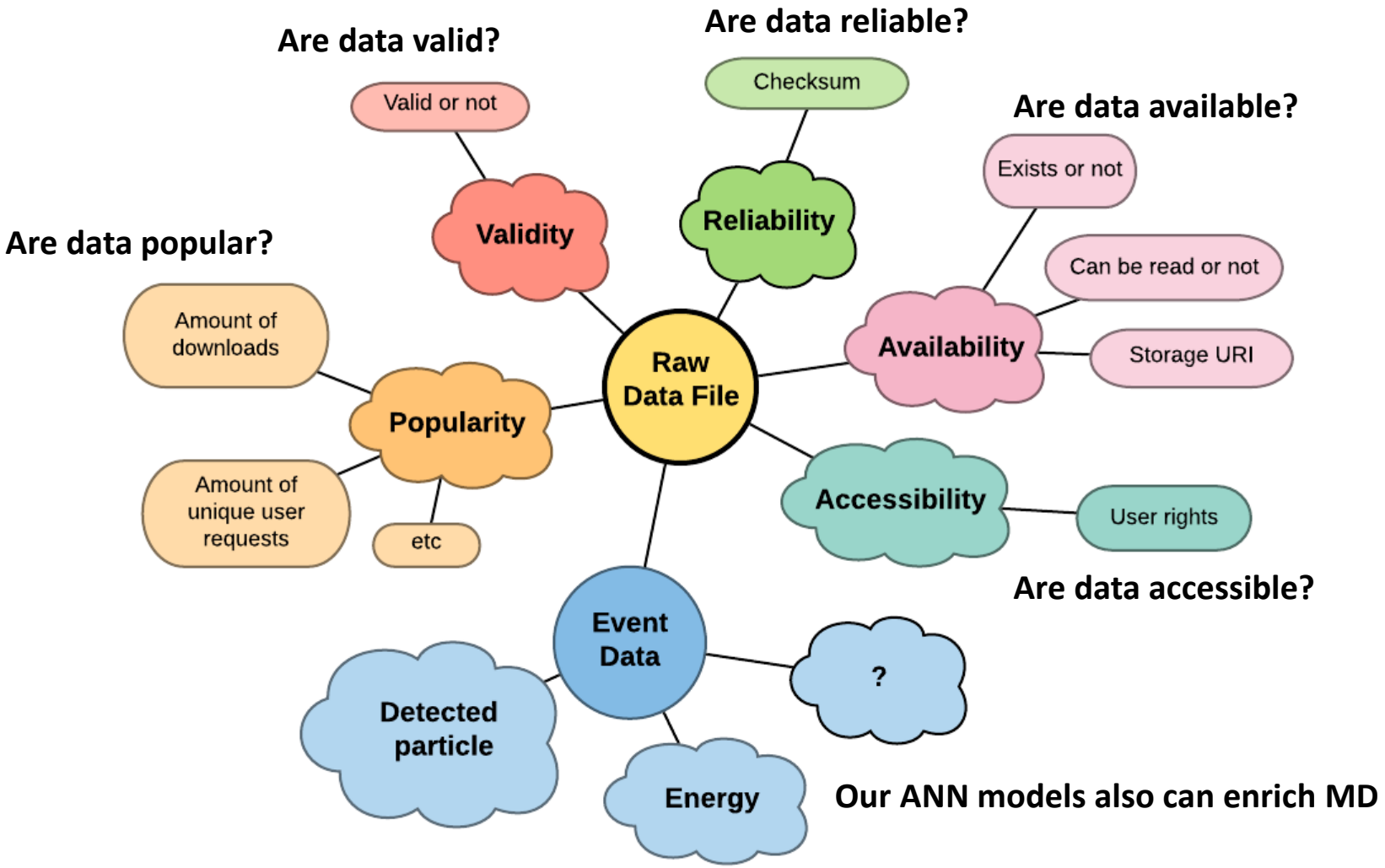
Metadata hidden in TAIGA Raw Data

Available From Raw Data Files

Facility-specific MD for an Event



Derived Metadata of TAIGA Raw Data



How We Can Extract Metadata from TAIGA Raw Data

Framework


```
FD 3F C5 5C 20 9D
B1 A2 3A 4D A8 B9
3F 45 04 6D DE E2
89 EC 67 71 74 14
09 80 F7 80 D0 A1
95 B8 95 66 F8 2A
53 8B 57 3D 8F F7
7F 44 69 20 C3 9A
```

TAIGA raw data

```
meta:
id: hiscore
title: HiSCORE
seq:
- id: packages
type: package
repeat: eos
types:
package:
seq:
- id: hdr
type: header
..
```

TAIGA format specification in YAML

```
...
vector<hiscore_t::package_t*>* packages =
hiscore.packages();
vector<hiscore_t::package_t*>::iterator it =
packages->begin();

for (it; it != packages->end(); ++it) {
  hiscore_t::package_t* package =
  (hiscore_t::package_t*)*it;
  hiscore_t::header_t* header = package->hdr();
  ...
}
...
```

C++ source code auto-generated by *Kaitai Struct*

Framework

Implemented part

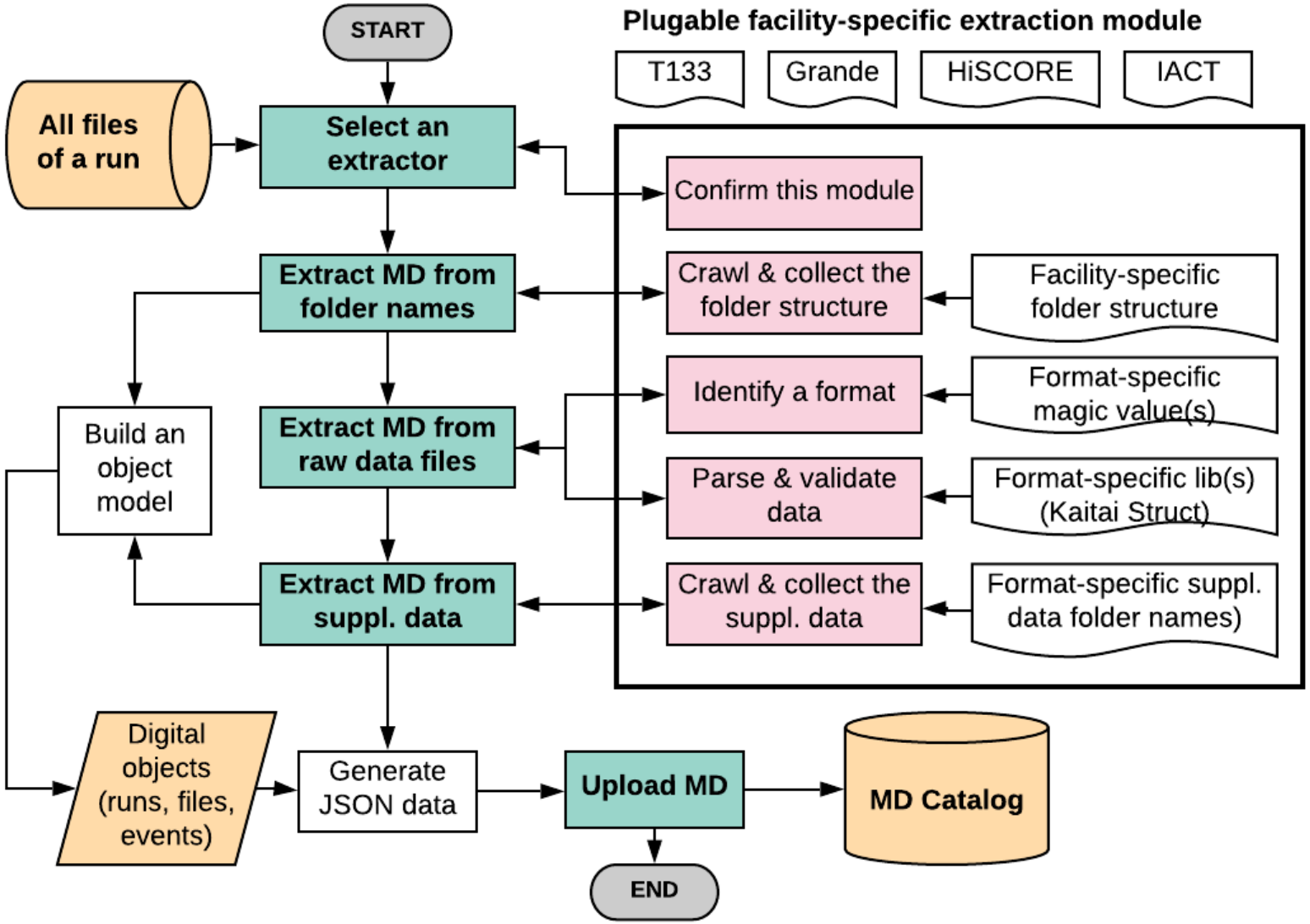
Should be implemented

Developed
5 specification for *TAIGA* file formats
- T133
- Grande
- TREX
- HiSCORE
- IACT

Generated
5 libraries for reading *TAIGA* raw data
- C/C++
- Python
- Java
- etc.

Tested on real *TAIGA* raw data
- T133, Grande, and TREX — **89K** files
- HiSCORE and IACT — **120K** files

MD Extractor Workflow



Implementation Requirements

- Micro-service architecture
- REST API
- Extensible architecture (facility-specific add-ons)

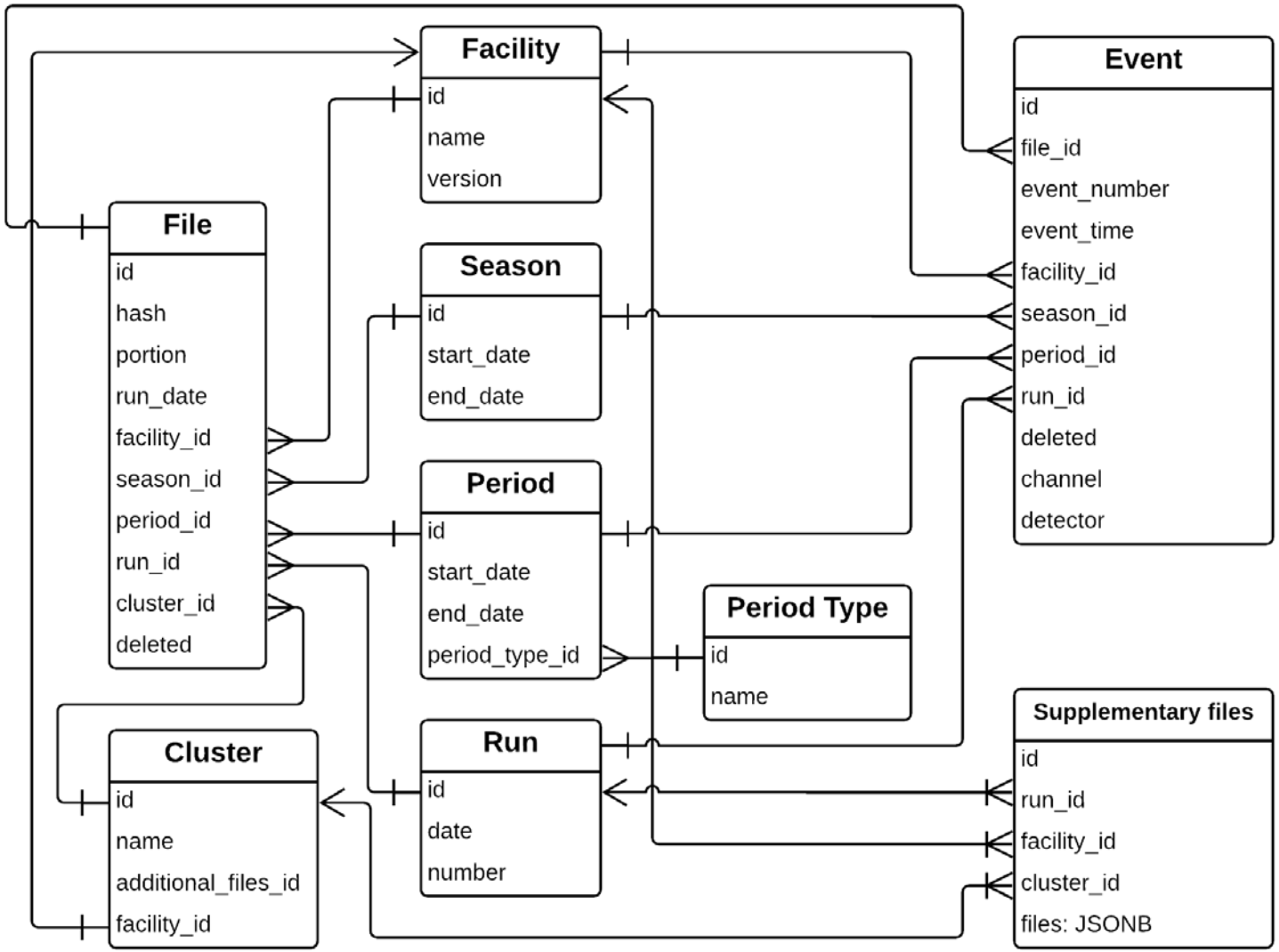
Where to We Can Extract Metadata from TAIGA Raw Data

Possible queries to the MD Catalog

- **GET data WHERE *time* ==**
  - **range** = time between start and end (less than a night)
  - **run** = a specified run | a calibration run
  - **night** = a specified date
  - **moonless month** = a period of time (not calendar month)
  - **summer** = a summer period of time

- **GET data WHERE *equipment* ==**
  - **facility** = a specified facility
  - **cluster** = a specified cluster (station) of a facility

Draft of the Metadata Catalog Schema



Conclusion

- **Two aspects of hidden MD**
  - Time
  - Equipment
- **Two parts of hidden MD**
  - Common attributes
  - Facility-specific attributes
- **Two kinds of user requests to MD**
  - Time
  - Equipment
- **Two parts of MD extractor’s architecture**
  - Unified workflow for MD extraction
  - Facility-specific extraction modules (add-ons)

Further Work

WHEN

- **Unify** the terminology: • “*task*” **vs** “*run*” • “*facility*” **vs** “*instrument*” • “*cluster*” **vs** “*station*” • **etc**
- **Define** a list of user requests to the MD catalog
- **Clarify** our understanding
  - *Is a channel or a detector also a digital object?*
  - *Which of facility-specific MD are needed?*
  - *Which of derived MD are needed?*
- **Understand** how to integrate the developing catalog with KCDC metadata

THEN

- **Design** the architecture and object model of the TAIGA MD extractor
- **Deploy** the MD catalog
- **Complete** the access API to the MD catalog
- **Implement** the TAIGA MD extractor

Thanks