

Automatic Extraction of Descriptive Metadata to Promote the Usage of RDM Tools

NFDI4Ing Conference 2022
2022-10-27

Benedikt Heinrichs



This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).



Introduction

- There is an ongoing task to transform research data to FAIR Digital Objects (FDOs)
- These FAIR Digital Objects contain
 - The Digital Object (research data)
 - Metadata about the Digital Object
 - Some Service Interfaces
 - A Persistent Identifier
- The big point we focus on here today is metadata

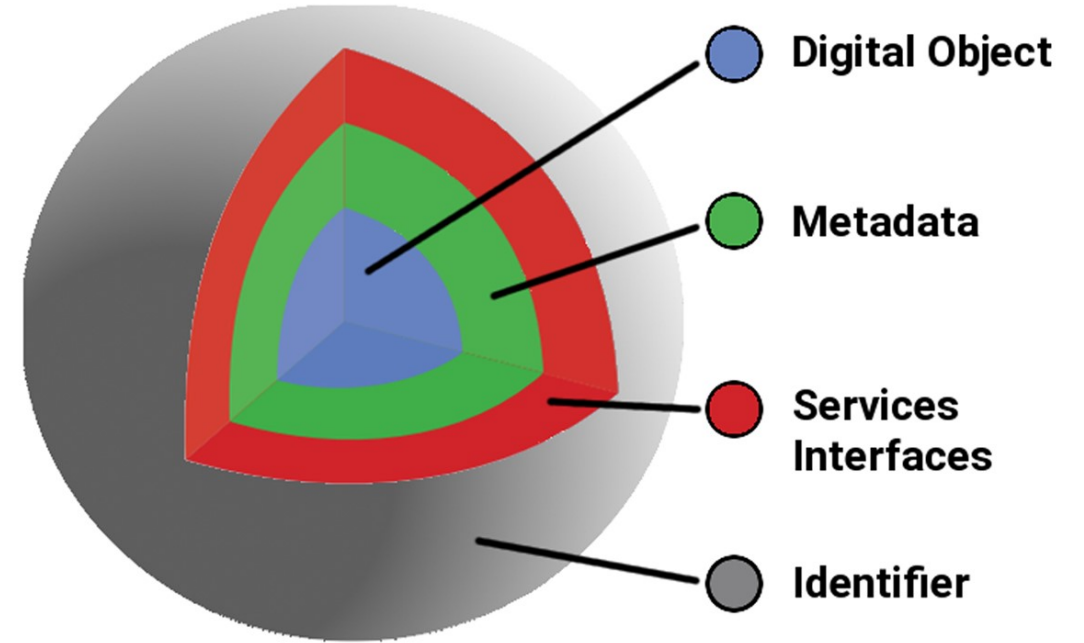


Image from "Digital Objects – FAIR Digital Objects: Which Services Are Required?", located at: <https://datascience.codata.org/articles/10.5334/dsj-2020-015/>

Metadata

- Why Metadata?

- Describing the information surrounding the generation of a research item
- Example: Describing a research experiment, the time it took place, etc.

- What Metadata?

- Administrative: e.g. location or rights
- Structural: provenance information
- Descriptive: who, when or what
- Formulated in RDF using ontologies and validated by SHACL

- How Metadata?

- Manually input
- Automatically generated during an experiment

Creator *	<input type="text" value="Benedikt Heinrichs"/>	✓	+
Title *	<input type="text" value="IC3K 2020 Poster"/>	✓	+
Production Date *	<input type="text" value="Wednesday, September 9, 2020"/>	✓	+
Subject Area	<input type="text" value="Informatik"/>	▼	+
Resource	<input type="text" value="Text"/>	▼	+
Rights	<input type="text"/>		+
Rightsholder	<input type="text"/>		+


Manual Metadata Input

- Administrative and Structural are usually fairly simple to automatically determine
 - e.g. a platform which manages metadata should be able to be aware of these types of metadata
- Descriptive metadata, however, currently mostly needs to be entered manually
 - This is usually a tedious and time-consuming task
 - A goal of RDM is to make the research process easier and not to create additional hurdles
 - Thankfully, the research data itself a lot of the time brings a subset of the necessary descriptive metadata with itself

Creator *	<input type="text" value="Benedikt Heinrichs"/>	✓	+
Title *	<input type="text" value="IC3K 2020 Poster"/>	✓	+
Production Date *	<input type="text" value="Wednesday, September 9, 2020"/>	✓	+
Subject Area	<input type="text" value="Informatik"/>		+
Resource	<input type="text" value="Text"/>		+
Rights	<input type="text"/>		+
Rightsholder	<input type="text"/>		+

Manual Metadata Input

- Administrative and Structural are usually fairly simple to automatically determine
 - e.g. a platform which manages metadata should be able to be aware of these types of metadata
- Descriptive metadata, however, currently mostly needs to be entered manually
 - This is usually a tedious and time-consuming task
 - A goal of RDM is to make the research process easier and not to create additional hurdles
 - Thankfully, the research data itself a lot of the time brings a subset of the necessary descriptive metadata with itself

Creator *	Benedikt Heinrichs	✓	+
Title *	IC3K 2020 Poster	✓	+
Production Date *	 Wednesday, September 9, 2020	✓	+
Subject Area	Informatik		+
Resource	Text		+
Rights			+
Rightsholder			+

Motivation for Automatic Metadata Extraction

- We want to know what our research data is about
- We want to provide the most detailed information about the content of our research data
- We want to spend only as much time as is necessary to input values into forms
- Proposition:



Result Motivation for Automatic Metadata Extraction



```
47      image:mode "RGB" ;
48      ebucore:hasFormat "JPEG" ;
49      ebucore:height "425" ;
50      ebucore:width "640" .
51
52      imageobject:apple rdfs:label "apple" ;
53      imageobject:count "6" .
54
55      imageobject:orange rdfs:label "orange" ;
56      imageobject:count "7" .
```

```
TestText.txt - Editor
Datei  Bearbeiten  Ansicht

Benedikt and Amin work in the same office.
David is eavesdropping their conversation.
David can also speak chinese and knows the characters '性格'.
```



```
68  <http://pikes.fbk.eu/#Amin> a <http://www.newsreader-project.eu/ontologies/PERSON> .
69
70  <http://pikes.fbk.eu/#Benedikt> a <http://www.newsreader-project.eu/ontologies/PERSON> .
71
72  <http://pikes.fbk.eu/#eavesdropping> rdfs:seeAlso <http://dbpedia.org/resource/Eavesdropping> .
```

Prepare the deployment for production server

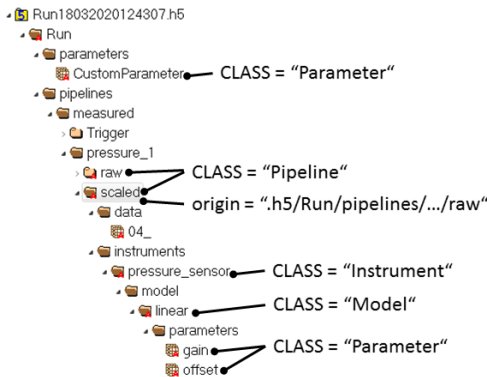
Refer to the product ticket for the description of the ticket.

- ☐ Update AutoSPInstaller.xml for installation
 - ☐ Check if there need to be separate SP Installer xmls for the Farm join since there is only one farm.
 - ☒ Check if we need a separate file. Set the secret keys and replace it with Consul values into the deployment script.
- ☒ Detect why the error during installation "Previously installed Office 2019" (PreReqCheck) occurred
- ☒ Parameter for the Installation => For Produktiv vs Dev
 - ☐ Create a parameter to filter (blacklist) steps
 - ☐ Add a json file with default values for config params, also secrets, check if this solves #322
- ☐ Make sure the production deployment is working
- ☒ Put CoScnE DB creation into its own step and otherwise throw 3.02 out

Topic/500-productionDeployment



```
87
88  <http://pikes.fbk.eu/#installation> a <http://dbpedia.org/class/yago/Initiation107453195> ;
89      rdfs:seeAlso <http://dbpedia.org/resource/Installation_(computer_programs)> .
90
91  <http://pikes.fbk.eu/#json> rdfs:seeAlso <http://dbpedia.org/resource/JSON> .
92
```

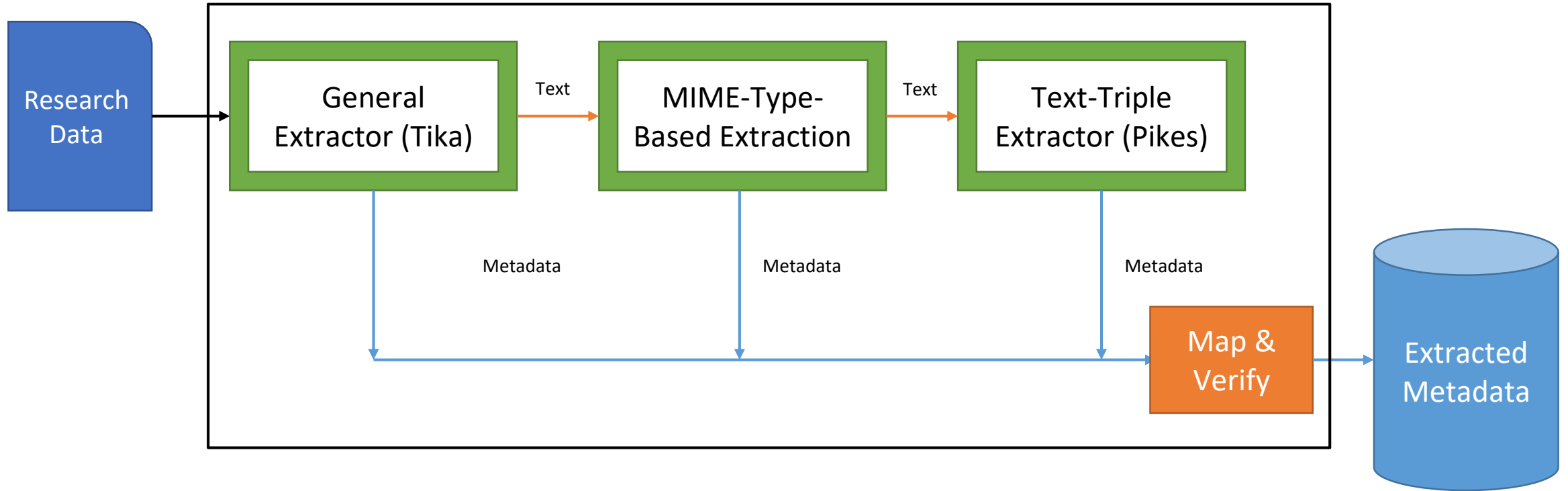


```
<https://hdl.handle.net/21.11102/
a ns3:Pipeline,
  dcat:Catalog ;
ns5:pipelineVersion "1.0" ;
ns3:origin "this" ;
ns3:units "volts" ;
ns3:variable "voltage" ;
dct:identifier "Run1803202012
dcat:catalog <https://hdl.han
<https://hdl.handle.net/2
```

HDF5 File Structure

Metadata Extraction Pipeline

Metadata Extraction



Heinrichs, B. ; Politze, M.

Moving Towards a General Metadata Extraction Solution for Research Data with State-of-the-Art Methods

12th International Conference on Knowledge Discovery and Information Retrieval, KDIR 2020, online, 2 Nov 2020 - 4 Nov 2020

Metadata Extraction Application – Features

- Dynamic Configuration
 - You can configure every extractor which shall be called and specify certain environment variables
- Registration of custom extractors
 - Once implemented, an extractor will listen to its registration method which can e.g. listen to certain MIME-Types like “image/png”
 - Custom extractors can be excluded from the default configuration, so that specific use cases can be proposed without impacting everything else
- Highly extendable
 - By being open source, this application is easily extendable to different use cases
 - <https://git.rwth-aachen.de/coscine/research/metadataextractor>

Metadata Extraction Service

Metadata Extractor API 0.1.1

[Base URL: /]
[/swagger.json](#)

This API extracts RDF triples from files

default

Default namespace

^

POST /

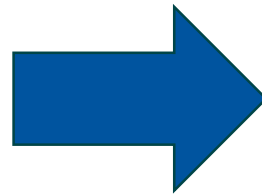
^

Parameters

Try it out

Name	Description
identifier string (formData)	File Identifier
	<input type="text" value="identifier"/>
config string (formData)	Object defining the utilized configuration (try "/defaultConfig" to get the structure)
	<input type="text" value="config"/>
creation_date string (formData)	Creation Date (Time) (e.g. "2022-09-15T09:27:17.3550000+02:00")
	<input type="text" value="creation_date"/>
modification_date string (formData)	Modification Date (Time) (e.g. "2022-09-15T09:27:17.3550000+02:00")
	<input type="text" value="modification_date"/>
file ★ required file (formData)	<input type="button" value="Datei auswählen"/> Keine ausgewählt

Example Results – Object Detection



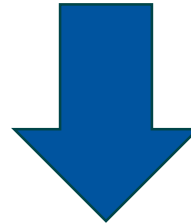
47		<code>image:mode "RGB" ;</code>
48		<code>ebucore:hasFormat "JPEG" ;</code>
49		<code>ebucore:height "425" ;</code>
50		<code>ebucore:width "640" .</code>
51		
52		<code>imageobject:apple rdfs:label "apple" ;</code>
53		<code>imageobject:count "6" .</code>
54		
55		<code>imageobject:orange rdfs:label "orange" ;</code>
56		<code>imageobject:count "7" .</code>

Example Results – Text

TestText.txt - Editor

Datei Bearbeiten Ansicht

Benedikt and Amin work in the same office.
David is eavesdropping their conversation.
David can also speak chinese and knows the characters '性格'.



```
68 <http://pikes.fbk.eu/#Amin> a <http://www.newsreader-project.eu/ontologies/PERSON> .
69
70 <http://pikes.fbk.eu/#Benedikt> a <http://www.newsreader-project.eu/ontologies/PERSON> .
71
72 <http://pikes.fbk.eu/#eavesdropping> rdfs:seeAlso <http://dbpedia.org/resource/Eavesdropping> .
```

Example Results – Image to Text

Prepare the deployment for production server

Refer to the product ticket for the description of the ticket.

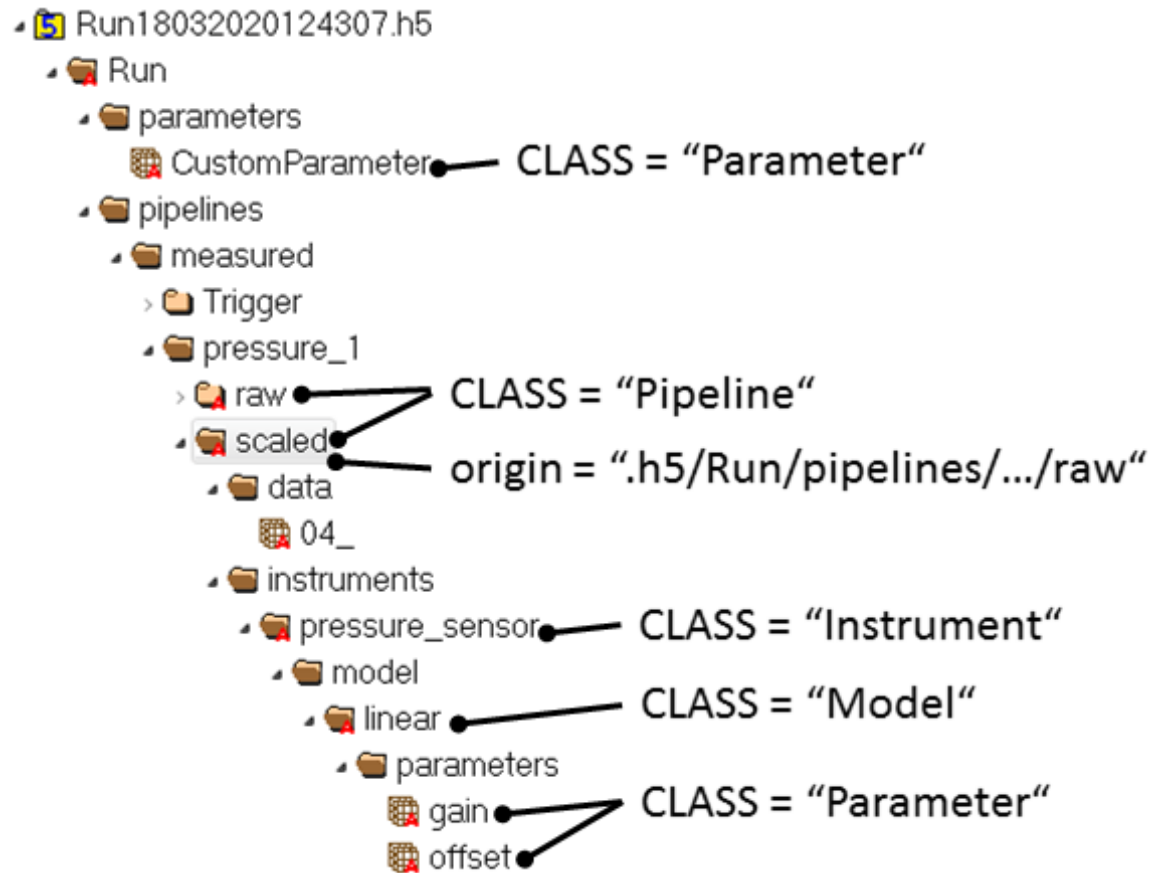
- ☐ Update AutoSPInstaller.xml for installation
 - ☐ Check if there need to be seperate SP Installer xmls for the Farm join since there is only one farm.
 - ☒ Check if we need a separate file. Set the secret keys and replace it with Consul values into the deployment script.
- ☒ Detect why the error during installation "Previously installed Office 2019" (PreReqCheck) occurred
- ☒ Parameter for the Installation => For Produktiv vs Dev
 - Create a parameter to filter (blacklist) steps
 - Add a json file with default values for config params, also secrets, check if this solves #322
- ☐ Make sure the production deployment is working
- ☒ Put CoScInE DB creation into its own step and otherwise throw 3.02 out

Topic/600-productionDeployment



```
87
88   <http://pikes.fbk.eu/#installation> a <http://dbpedia.org/class/yago/Initiation107453195> ;
89   rdfs:seeAlso <http://dbpedia.org/resource/Installation\_\(computer\_programs\)> .
90
91   <http://pikes.fbk.eu/#json> rdfs:seeAlso <http://dbpedia.org/resource/JSON> .
92
```

Example Results – Real Research Data Example with HDF5



HDF5 File Structure



```
<https://hdl.handle.net/21.11102/1
  a ns3:Pipeline,
    dcat:Catalog ;
  ns5:pipelineVersion "1.0" ;
  ns3:origin "this" ;
  ns3:units "volts" ;
  ns3:variable "voltage" ;
  dct:identifier "Run18032020124307.h5" ;
  dcat:catalog <https://hdl.handle.net/21.11102/1> ;
  <https://hdl.handle.net/21.11102/1>
```

Heinrichs, B. P. A. ; Preuß, N. ; Politze, M. ; Müller, M. S. ; ; et al

Automatic General Metadata Extraction and Mapping in an HDF5 Use-case

Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management

Metadata Extraction Service – Usage

Open Source

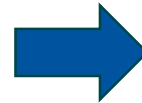
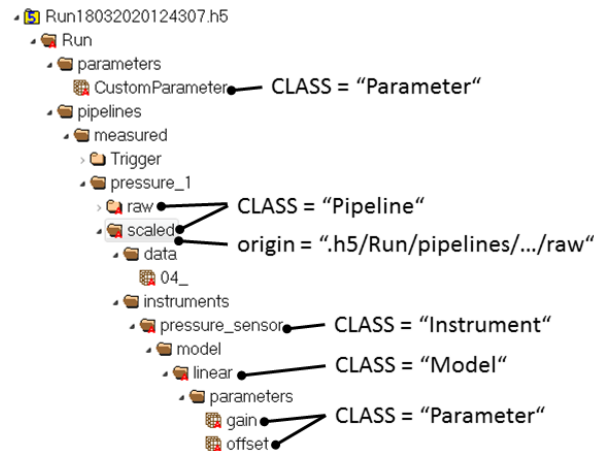
- Git Repo (Python Code):
<https://git.rwth-aachen.de/coscine/research/metadataextractor>
- Docker Image:
registry.git.rwth-aachen.de/coscine/research/metadataextractor:latest
- Demo:
<https://metadataextractor.otc.coscine.dev/>
- Start adding your own extractor now by using Gitpod:
<https://gitpod.io/#https://git.rwth-aachen.de/coscine/research/metadataextractor>

Metadata Extraction Service – Use Cases (Future Work)

- Inclusion in research data management systems like Coscine
 - Automatically extract the metadata for research data based on a given configuration
 - Fill the metadata form automatically based on a templating engine
 - Make use of the extracted metadata in applications like “search”
- Utilize the extracted metadata to determine the similarity between research data when the MIME-Type is different (e.g. image with text vs. text file)
 - Making use of <https://git.rwth-aachen.de/coscine/research/semanticsimilarity>
- Improve the performance and make it better scale against big research data

Conclusion

- Today, I showed a look into my proposed solution of metadata extraction
- It is a pipeline which takes research data and tries to describe the content as metadata
- The usage has been demonstrated on different examples with a real-life use case as well



```
<https://hdl.handle.net/21.11102/1
a ns3:Pipeline,
  dcat:Catalog ;
ns5:pipelineVersion "1.0" ;
ns3:origin "this" ;
ns3:units "volts" ;
ns3:variable "voltage" ;
dct:identifier "Run18032020124307" ;
dcat:catalog <https://hdl.handle.net/21.11102/1>
<https://hdl.handle.net/21.11102/1>
```

HDF5 File Structure

- Future work is being done to utilize this in real life applications

**Thank you
for your attention!**