Data Processing: Ingest Data Model

Matthew Templeton, Kelly Lockhart, and the ADS Team

ADS Users Group Meeting, 15-16 Nov. 2021

Current data ingest

- Ingest is publisher- and format-based
- Output to formatted text files
- Complex file naming/directory heirarchy for files (including records with multiple sources)
- Separate processing for fulltext, references

Title: Agile Methodologies in Teams with Highly Creative and Autonomous Members

Authors: Blanco-Cuaresma, S.; Accomazzi, A.; Kurtz, M. J.; Henneken, E. A.; Grant, C. S.; Thompson, D. M.; Chyla, R.; McDonald, S.; Shapurian, G.; Hostetler, T. W.; Templeton, M. R.; Lockhart, K. E.; Bukovi, K.; Rapport, N.

Affiliation: AA(Harvard-Smithsonian Center for Astrophysics, HEAD, Cambridge, MA, 02138, USA <EMAIL>sblancocuaresma@cfa.harvard.edu</EMAIL> <ID system="ORCID">0000-0002-1584-0171</ID>) AB(Harvard-Smithsonian Center for Astrophysics, HEAD, Cambridge, MA, 02138, USA <EMAIL>aaccomazzi@cfa.harvard.edu</EMAIL> <ID system="ORCID">0000-0002-4110-3511</ID>) [...]

Journal: Astronomical Data Analysis Software and Systems XXIX. ASP Conference Series, Vol. 527, proceedings of a conference held (6—10 October 2019) at the Martini Plaza, Groningen, the Netherlands. Edited by Roberto Pizzo, Erik R. Deul, Jan David Mol, Jelle de Plaa, and Harro Verkouter. San Francisco: Astronomical Society of the Pacific, 2020, p.505

Publication Date: 00/2020

Origin: ASP

Bibliographic Code: 2020ASPC..527..505B

Abstract

The Agile manifesto encourages us to value individuals and interactions over processes and tools, while Scrum, the most adopted Agile development methodology, is essentially based on roles, events, artifacts, and the rules that bind them together (i.e., processes). Moreover, it is generally proclaimed that whenever a Scrum project does not succeed, the reason is because Scrum was not implemented correctly

ADS Ingest Data Model:JSON Schema

- Object- and content-oriented approach: keep related information together
- Limit / eliminate reinterpretation of publisher data at parse time
- Keep processing history and provenance as part of each record
- Allow schema revision/expansion as needed
- Validation: software to validate at parse time

```
{
  "$schema": "http://json-schema.org/draft-07/schema#",
  "title": "Document",
  "description": "Data schema for importing metadata into ADS pipeline",
  "type": "object",
  "properties": {
    "recordData": {
      "$ref": "./RecordData.json"
    },
    "relatedTo": {
      "$ref": "./RelatedTo.json"
    },
    "editorialHistory": {
      "$ref": "./EdHist.json"
    },
    "pubDate": {
      "$ref": "./PubDates.json"
    },
    "publication": {
      "$ref": "./Publication.json"
    },
    "persistentIDs": {
      "type": "array",
      "description": "Array of PersistentID.json objects",
      "items": {
        "$ref": "./PersistentID.json"
      }
    },
    "publisherIDs": {
      "type": "array",
      "description": "Array of PublisherID.json objects",
      "items": {
        "$ref": "./PublisherID.json"
      }
    },
    "pagination": {
      "$ref": "./Pagination.json"
    },
```

```
{
  "title": "RecordData",
  "description": "Data schema for importing metadata into ADS pipeline",
  "type": "object",
  "properties": {
    "createdTime": {
      "description": "Timestamp for when the metadata was harvested (e.g. file create
d timestamp)",
      "type": "string"
    },
    "parsedTime": {
      "description": "Timestamp for when parsing commenced.",
      "type": "string"
    },
    "loadType": {
      "description": "ENUM: do we get this from a file or from a URL (or other?)",
      "$comment": "This list may need to be expanded.",
      "type": "string",
      "enum": [
        "fromFile",
        "fromURL"
      ]
    },
    "loadFormat": {
      "description": "ENUM:",
      "$comment": "This list may need to be expanded.",
      "type": "string",
      "enum": [
        "JATS",
        "OtherXML",
        "HTML",
        "Text"
      ]
    },
    "loadLocation": {
      "description": "If loadtype is fromFile, path to file; if fromURL, it's a URL",
      "type": "string"
    },
```

```
{
  "recordData": {
    "createdTime": "2021-08-31Z00:00:00",
    "parsedTime": "2021-09-15Z12:00:00",
    "loadType": "fromFile",
    "loadFormat": "JATS",
    "loadLocation": "[...]/data/A+A/A+A652/abstracts/aa37735-20.xml",
    "recordOrigin": "Publisher"
  },
  "pubDate": {
    "printDate": "2021-08-27",
    "electrDate": "2021-08-27"
  },
  "title": {
    "textEnglish": "Diagnostic capabilities of spectropolarimetric observations for understanding solar phenomena"
  },
  "subtitle": "I. Zeeman-sensitive photospheric lines",
  "keywords": [
    {
      "keySystem": "Astronomy",
      "keyString": "Sun: magnetic fields, techniques: polarimetric, atomic data, Sun: photosphere, radiative transfer"
    }
  ],
  "authors": [
    {
      "name": {
        "surname": "Quintero Noda",
        "given-name": "C."
      },
      "attrib": {
          "email": "carlos.quintero@iac.es",
          "orcid": "0000-0001-9218-3139"
      },
      "affiliation": [
        {
          "affPubRaw": "<label>1</label> <addr-line> <institution>Rosseland Centre for Solar Physics, University of Oslo</institution>,
<named-content content-type=\"postbox\">PO Box 1029</named-content>, <named-content content-type=\"city\">Blindern</named-content> <name
d-content content-type=\"postcode\">0315</named-content> <named-content content-type=\"state\">Oslo</named-content>, <country>Norway</co
untry> </addr-line>",
          "affPubRaw": "<label>2</label> <addr-line> <institution>Institute of Theoretical Astrophysics, University of Oslo</institution
>, <named-content content-type=\"postbox\">PO Box 1029</named-content>, <named-content content-type=\"city\">Blindern</named-content> <n
amed-content content-type=\"postcode\">0315</named-content> <named-content content-type=\"state\">Oslo</named-content>, <country>Norway<
/country> </addr-line>",
```

Next steps: processing infrastructure

- Generalized parsers for content delivery formats
  - Adsabs-pyingest will be a model but we're redesigning the process
- Parsing: file/structure validation system
  - Currently testing with python: jsonschema package
- Integrated ingest tracking and monitoring
  - Alert curators of issues at parse time (rather than index time)
  - Slack notifications, curator dashboard(?)
- Develop Storage Data Model (superset of ingest model), new record identifier system

Validation:


```
{
  "recordData": {
    "createdTime": "2021-08-30Z12:00:00",
    "parsedTime": "2021-10-02Z18:50:00",
    "loadType": "fromFile",
    "loadFormat": "Text",
    "loadLocation": "./real_data/14500.gcn3",
    "recordOrigin": "Publisher"
  },
  "pubDate": {
    "electrDate": "2013-04-29"
  },
  "title": {
    "textEnglish": "GRB 100728A: GROND host detection and X-shooter redshift"
  },
  "authors": [
    {
      "name": {
        "surname": "Kruehler",
        "given-name": "T."
      },
      "affiliation": [
        {
          "affPubRaw": "DARK"
        }
      ]
    },
    {
      "name": {
        "surname": "Greiner",
        "given-name": "J."
      },
      "affiliation": [
        {
          "affPubRaw": "MPE"
        }
      ]
    },
```



```
  },
  "pubDate": {
    "electrDate": "2013-04-29"
  },
  "title": "GRB 100728A: GROND host detection and X-shooter redshift",
  "authors": [
    {
      "name": {
        "surname": "Kruehler",
        "given-name": "T."
      },
```


```
Now testing: ./gcnc_bad.json

File ./gcnc_bad.json failed:
        Error: 'GRB 100728A: GROND host detection and X-shooter redshift' is not
 of type 'object'

Failed validating 'type' in schema['properties']['title']:
    {'properties': {'langNative': {'type': 'string'},
                    'textEnglish': {'type': 'string'},
                    'textNative': {'type': 'string'}},
     'type': 'object'}

On instance['title']:
    'GRB 100728A: GROND host detection and X-shooter redshift'
File ./gcnc_bad.json failed
```