eawag aquatic research

Research Data Platform

ckan FOR RESEARCH DATA MANAGEMENT AT EAWAG

— A REPORT FROM THE TRENCHES —

CKANCon 2016

Harald von Waldow <harald.vonwaldow@eawag.ch> — 2016-10-04

eawag aquatic research

Research Data Platform

THE *PROJECT*

ESTABLISH AN *INSTITUTION-WIDE*, *CENTRAL* RESEARCH DATA REPOSITORY

- Compliance with Guidelines for Good Scientific Practice (e.g., reproducibility, retention obligation)

- Archival of unique unrepeatable measurements (≈ everything in the environment)

- Data discovery and collaboration across working groups, departments (and institutions).

eawag aquatic research | Research Data Platform

Eawag is concerned with concepts and technologies for dealing sustainably with water bodies and with water as a resource.

- ~ 500 staff in research, teaching and consulting
- 12 research departments
- ~ 70 research groups

eawag aquatic research ooo { Research Data Platform

A VERY HETEROGENEOUS DATA-MIXTURE

e.g.

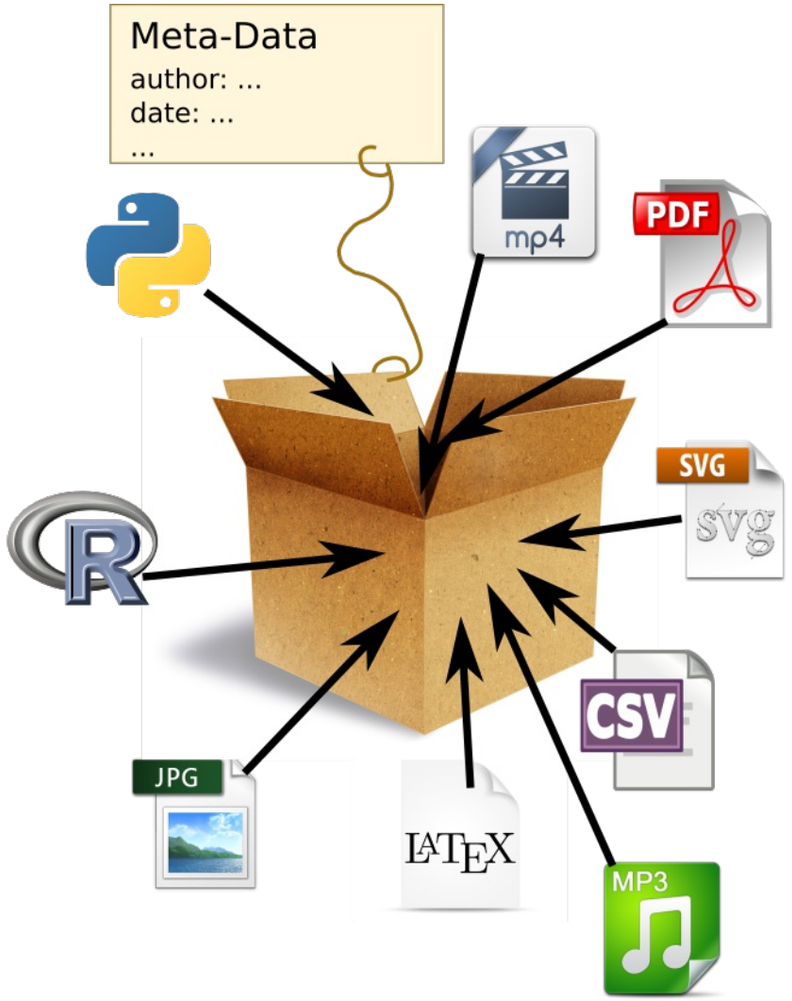
- high performance mass-spectrometry
- genomics & transcriptomics
- hydrologic data
- species counts
- urban drainage sensor networks
- surveys & questionnaires

- file-sizes: kB - 100s of GB
- #files per package: 1 - 100s
- analysis scripts
- CSV, PDF, baroque Excel files
- binary data-logger output
- photographs, movies
- ...

eawag aquatic research | Research Data Platform

CHARACTERISTICS OF PACKAGES (= "*DATASETS*")

- publication packages
- raw data
- the workgroup "stock-data"



TIME HORIZONS FOR STORAGE

- mid-term: ~ 10 years
- long-term: ~ 100 years

eawag aquatic research

Research Data Platform

Lib4RI

OSPER

opendata.swiss

zenodo

Active Directory

BAR

custom | OAI-PMH | CKAN native

extensions

Celery tasks

Research Data Platform

Open Data Mirror

LDAP

REST API

Web-UI

extensions

extensions

backup, versioning

custom interface

ETH LTS

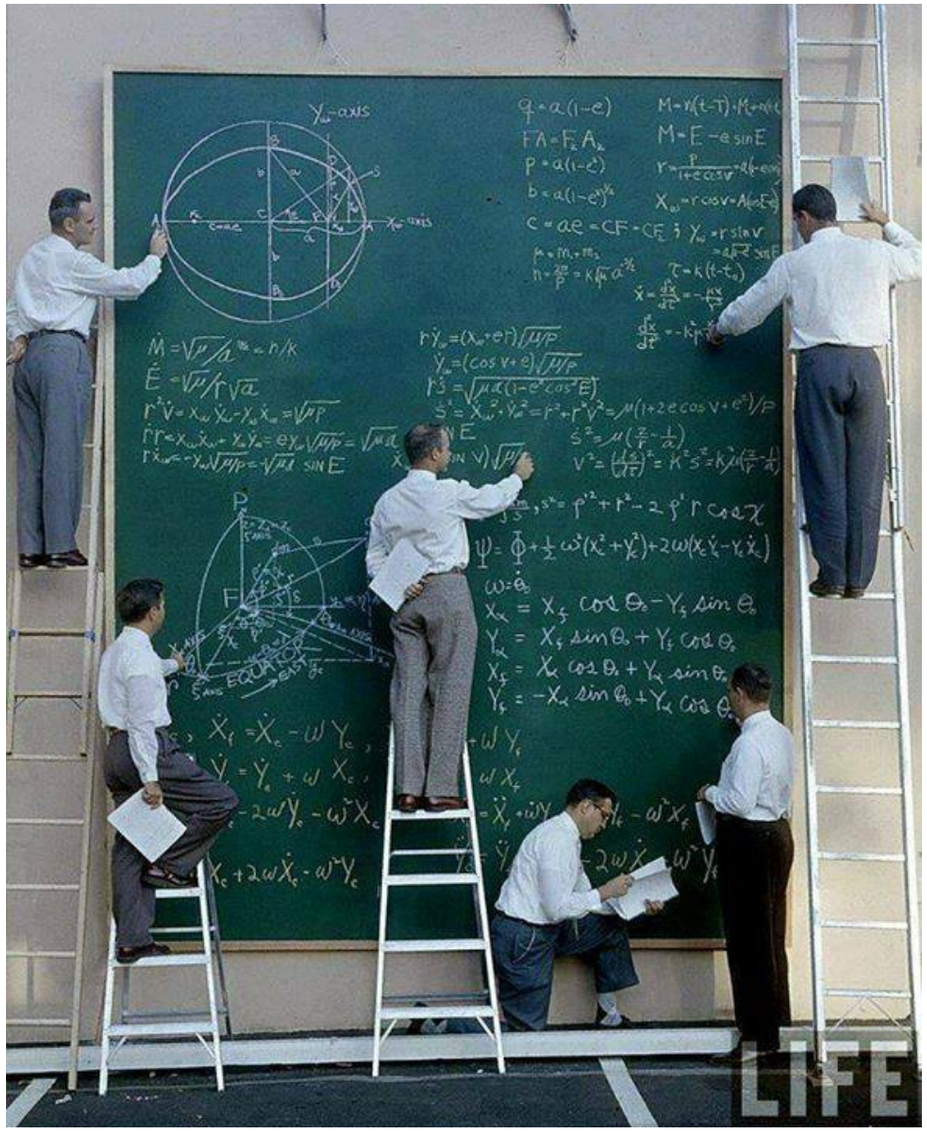
eawag aquatic research

eawag aquatic research

Research Data Platform

THE *CHALLENGE*

eawag aquatic research ooo { Research Data Platform



result:

- well annotated
- good file formats
- complete
- consistent

submission process:

- quick
- simple
- intuitive

⇌ ...

eawag aquatic research

Research Data Platform

AMENDMENTS FOR

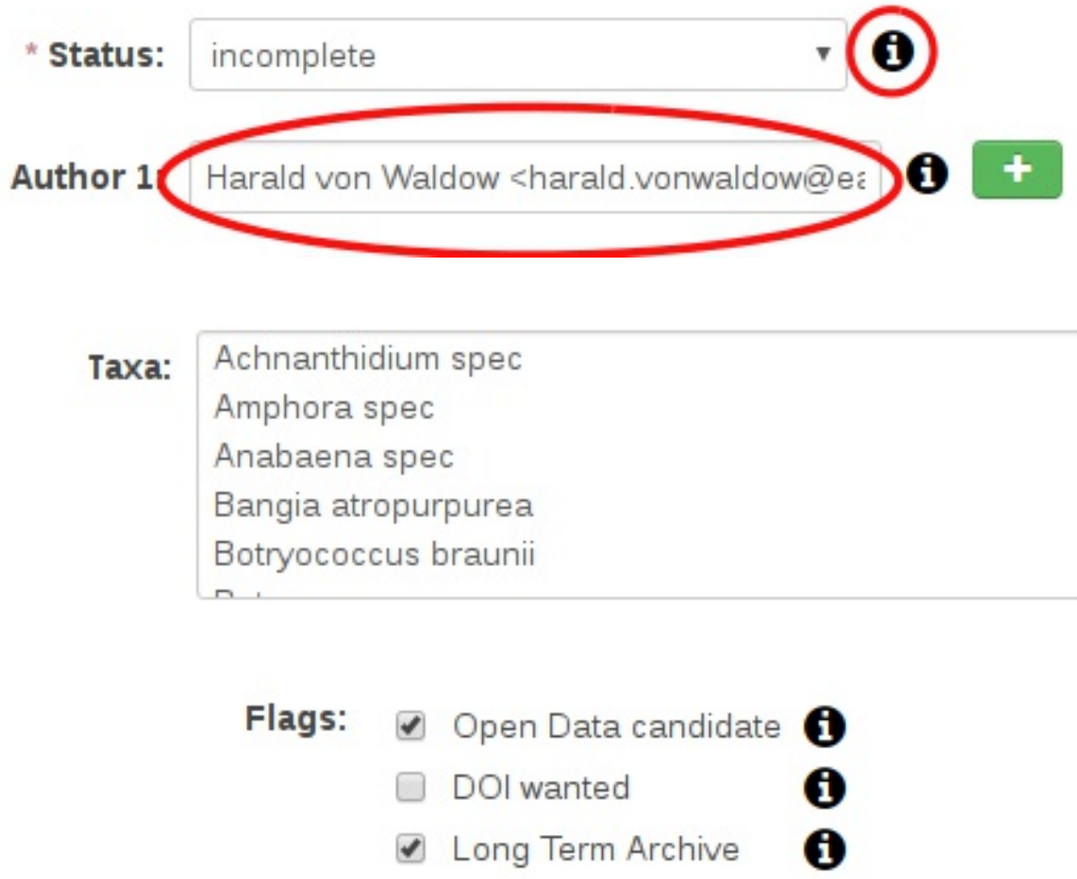
DATA SUBMISSION

eawag aquatic research | Research Data Platform

CKANEXT-SCHEMING IS THE THING!

`small amendments`

- extensive help-modals
- (dynamic) default values
- ckanext-repeating.js [lcrnz-repeating.js]
- line-by-line textareas
- checkbox array with constraints



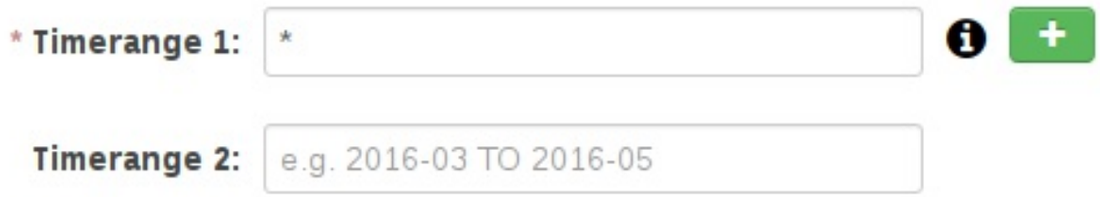
eaw_schema_conditional_multicheck.js

CKANEXT-SCHEMING IS STILL THE THING!

`more involved amendments`

Use the SOLR DateRange field

```
def vali_daterange(values)
validate_solr_daterange.py
```



Index JSON - store string

```
def eaw_schema_multiple_string_convert(typ)

@scheming_validator
def eaw_schema_multiple_choice(field, schema)

#IPackageController
def before_index(self, pkg_dict)
```

```
json2list_fields = [
  'substances',
  'variables',
  'systems'
  'timerange',
  'taxa',]
```

eawag aquatic research | Research Data Platform

CLIENT SIDE SCRIPTS

```
resup.py [-h] {put,get,list,del} ...
```

- Batch upload of resources to data package in CKAN.
- Batch download and deletion from data package in CKAN.

resup handles compression, creation of a tar-archive, checksumming, splitting of large files for upload, and re-assemblage of thusly split files upon download.

Resources to be downloaded or deleted can be specified by providing a regular expression to select resource names.

LARGE TODO-ITEM: GEOREFERENCING

- copy & paste coordinates
- accept lat/lon, lon/lat, CH1903, CH1903+, ...
- parse degrees minutes (seconds) into decimal degrees
- graphical selection of locations
- points and lines and polygons

Florian Mayer
florianm

ckanext-spatial/pull/93

eawag aquatic research ooo { Research Data Platform

LARGE TODO-ITEM: VERSIONING

package level: `http://.../dataset/history/<name>`

- Is this going to stay? (currently undocumented)
- Is there a rollback-extension?

resource level: `nothing implemented yet?`

Idea: Add resource_fields:

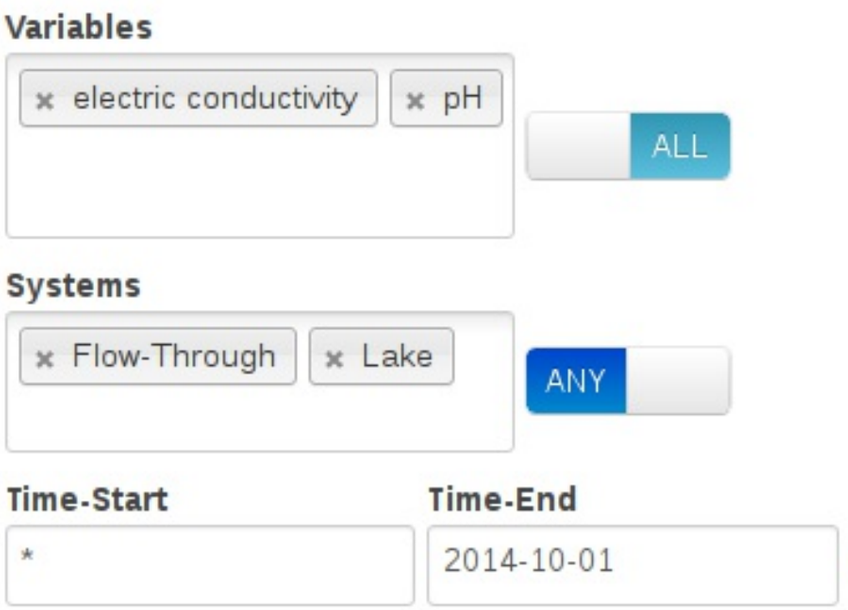
- supersedes: <resource_id>
- version: <maj.min.pat>

eawag aquatic research

Research Data Platform

AMENDMENTS FOR THE *SEARCH INTERFACE*

FIELD SPECIFIC SEARCH



- IPackageController:before_search
- parse custom search_params
- into SOLR search-string

```
def mk_field_queries(search_params):
    '''
    Customizes the fq-search-string so that query-terms
    referring to the same  (e.g. "example_field") are combined
    with logic operator taken from the value of OP_,
    e.g. "OP_example_field". Default for this operator is "AND" to
    keep compatibility. The other possible value is "OR".
    OP_ is removed from the querystring.
    '''
```

eawag aquatic research

Research Data Platform

A MODEST PROPOSAL

eawag aquatic research ooo { Research Data Platform

FUTURE CKAN DEVELOPMENT COULD

- integrate `ckanext-scheming` into core
- eliminate special (aka "core") fields
  (provide a default `scheming.json` instead.)
- streamline handling of structured schema items
- foster a "scheming-library" of plugins and snippets

QUESTIONS?