ESS-DIVE
Advanced Bootcamp

November 30th, 2020 10:00 - 11:00 am PDT
Which project or group are you representing today?

“Watershed Function SFA”

“NGEE-Arctic”

“ESS-DIVE”

“Brookhaven National Lab (NGEE projects)”

“Various: TES at PNNL, ESS CF Partners project, ...”
What is your role in your group?

- Principal Investigator (PI): 40%
- Data Manager: 40%
- Researcher: 20%
- Field Researcher
- Student
Which session are you most interested to learn more about today?

- Package Service API: 100%
- Customizable Data Portals
- Data Reporting Formats
Which of these topics are you familiar with?

- Package Service API
- Customizable Data Portals
- Data Reporting Formats
- None of the Above

100%
ESS-DIVE
Advanced Bootcamp
November 30th, 2020 10:00 - 11:00 am PDT

Madison Burrus, Rob Crystal-Ornelas, Fianna O’Brien
Zarine Kakalia, Emily Robles, Deb Agarwal, Charuleka Varadharajan, Shreyas Cholia, Valerie Hendrix, Joan Damerow, Hesham Elbashandy, Mario Melara, Makayla Shepherd, Maegen Simmonds, Cory Snavely, and Karen Whitenack
Advanced Bootcamp Overview

- **Programmatic** submissions and updates to data packages using the Package Service API
- **Customizable data portals** to organize your data collections
- How to use the ESS-DIVE community github for reviewing and commenting on the ESS-DIVE proposed metadata and data reporting formats

*This is an interactive webinar! We have demos, polls, and question breaks prepared.*
ESS-DIVE Package Service API
What is it?

We’ve discussed upload via the web interface. What if that becomes cumbersome?

Through **ESS-DIVE Package Service API**, organizations can write code to store, modify, and search data packages on ESS-DIVE.
Package Service Use Cases

When to use this

- Uploading **large data**
- **Programmatic** upload
- **Bulk data** upload
- **Reusing information** across data packages, such as location data.

Who should use this?

- Projects with a large number of data packages
- Teams wanting to duplicate package metadata
- Projects with consistent metadata standards
<table>
<thead>
<tr>
<th>Rank</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>R</td>
</tr>
<tr>
<td>2</td>
<td>Bash shell</td>
</tr>
<tr>
<td>3</td>
<td>I'm not familiar with coding</td>
</tr>
<tr>
<td>4</td>
<td>R</td>
</tr>
<tr>
<td>5</td>
<td>C++</td>
</tr>
</tbody>
</table>
Which of these terms are you familiar with?

- API
  - A: 50%
- REST API
  - B: 33%
- JSON-LD
  - C: 17%
How many data packages do anticipate submitting to ESS-DIVE in total by the end of 2021?

- < 5: 33%
- 5 - 10: 33%
- 10 - 20: 33%
- 20+: 33%
Package Service API Capabilities

● **Create & update data packages**
  ○ Update metadata
  ○ Add data files
  ○ Replace existing data files

● **Upload Large Data Files**

● **Explore existing package metadata**
  ○ Full metadata in single package search, including data files, with easy links to website

● **Package Service Documentation**
  ○ Easy to navigate tutorials
  ○ Troubleshooting guides
  ○ Helpful tips

[docs.ess-dive.lbl.gov](docs.ess-dive.lbl.gov)
Package Service API Upload Process

1. Review tutorial documentation for Package Service API
2. Get authentication token from ESS-DIVE
3. Create package metadata in JSON-LD
4. Upload Metadata & Data

Success! View package landing page on ESS-DIVE

Errors? Fix any errors found by validation using documentation
API Documentation Tutorial

To get started with the Package Service API, we recommend using ESS-DIVE’s API Tutorial.

This tutorial includes:

● All prerequisites needed to use the API
● Code examples of how to submit, update, and search data packages
● Complete examples in three different coding languages:
  ○ R
  ○ Python
  ○ Java
Accessible Coding

- Tutorial documentation gives complete walkthrough of API submission
- API metadata fields are equivalent to the fields in web upload form
- Quick learning curve for users of varying experience
## Metadata Crosswalk example

<table>
<thead>
<tr>
<th>ESS-DIVE Field</th>
<th>JSON-LD</th>
<th>DataCite 4.1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title</td>
<td>name</td>
<td>title</td>
</tr>
<tr>
<td>Alternative Identifiers</td>
<td>alternateName</td>
<td>alternateldIdentifiers</td>
</tr>
<tr>
<td>Abstract</td>
<td>description</td>
<td>description[@type=Abstract]</td>
</tr>
<tr>
<td>Keywords</td>
<td>keywords</td>
<td>subjects</td>
</tr>
<tr>
<td>Data Variables</td>
<td>variablesMeasured</td>
<td>subjects</td>
</tr>
<tr>
<td>Publication Date</td>
<td>datePublished</td>
<td>publicationYear</td>
</tr>
</tbody>
</table>

JSON-LD Metadata Standard

- **JSON object notation** is easier to read, understand & manipulate for humans than many other object notation styles.
- JSON-LD embedded in data package landing page allows harvesting by Google.
Create new data package

- You will be able to **validate and submit JSON-LD** against the interface.
- If the JSON-LD is invalid, details about the errors will be given.
- You may upload files up to **100 GB**.
API Submission Responses

```
"detail": "One or more fields raised validation errors.",
"errors": [
    "'datePublished' is a required property",
    "'provider' is a required property",
    "'editor' email' is a required property"
]
```

Error

```
"id": "ess-dive-c5ab2ee757aebe4c-20190927T203522703923",
"viewUrl": "https://data-sandbox.ess-dive.lbl.gov/view/
    ess-dive-c5ab2ee757aebe4c-20190927T203522703923",
"detail": "Data Package created successfully.",
"errors": null,
"dataset": {
    "name": "ESS-DIVE Webinar Demo",
    "description": "Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non"
```
Retrieve single data package metadata

- **Retrieve a single data package submitted by you with complete metadata**
  - View complete JSON-LD record
  - View list of associated uploaded data files

Upload a package via the web & retrieve it via API to **see an example of JSON-LD with your data**
Update existing data package

- **Metadata Updates**
  - Submit only the fields you want to update using JSON-LD

- **Data File Updates**
  - Upload additional files to an existing data package

```json
{
  "spatialCoverage": [
    {
      "description": "Site ID: S1 Bog Site name: S1 Bog, Marcell"
    },
    {
      "geo": [
        {
          "name": "Northwest",
          "latitude": 47.50285,
          "longitude": -93.48283
        },
        {
          "name": "Southeast",
          "latitude": 47.50285,
          "longitude": -93.48283
        }
      ]
    }
  ]
}```
Retrieve list of data package metadata

- **Retrieve a complete list** of data packages submitted by you with important metadata, including:
  - Title
  - URL
  - Public Visibility
  - Citation
Common Errors

- **JSON-LD validation errors**
  - Use tutorials, technical documentation, or packages examples to understand JSON-LD schema expectations.

- **Out of date token**
  - Tokens expire daily. Make sure token is up to date.
API Technical Documentation

For technical details about ESS-DIVE’s Package Service API, visit our **extended documentation**.

- This is hosted on a separate webpage from other ESS-DIVE documentation
- It’s best to use this resource when **troubleshooting errors**
Package Service API Summary

- Programmatically upload, update, and search data packages
- Reuse code for duplicated information across packages
- Upload large data files & perform bulk uploads
- Use tutorial documentation to get comfortable

Tutorial documentation: docs.ess-dive.lbl.gov
API documentation: api.ess-dive.lbl.gov
Customized Data Portals
What is a data portal?

A data portal is a collection of any ESS-DIVE data packages. Researchers, projects, labs, and even individuals can easily highlight and share their datasets and research topics using a data portal.
ESS-DIVE’s Data Search Interface can also be considered a data portal. In this data view, we can see every data package ever published in ESS-DIVE.
Data Portals: a unique data view

In a customizable data portal, we can select which data packages we view at once and shift the focus of ESS-DIVE’s default data view.
Key Portal Features

● Portals can be **branded**
  ○ All portals have **customizable** banners, logos, and URLs

● Provide a space for **data and research context**

● Contain portal specific **data metrics**

● Easily **share** your portals

[https://data.ess-dive.lbl.gov/portals/CDIAC](https://data.ess-dive.lbl.gov/portals/CDIAC)
Key Portal Features

- You can imagine a portal as a personalized website for your research that’s hosted by ESS-DIVE

- If you already have a website for your research, the portal link can be inserted into your website to allow easy access to your data

Portals are a powerful tool for connecting related datasets together
Portals give projects and groups an identity in ESS-DIVE

- Portals allow you to give your datasets scientific context beyond what’s capable with just your data package alone.
- Your portal can describe...
  - Your research group
  - The study site
  - The research question your addressing
Portals give projects and groups an identity in ESS-DIVE

In the CDIAC portal, we focus on the scientific purpose of this data collection
Data portal specific metrics

This metrics page is personalized to the data packages that you choose to add to your portal.
Data portal specific metrics

- Some additional metrics are:
  - Number of times your portal data has been cited
  - Number of times the portal data has been downloaded
  - Number of times your data package landing pages have been viewed

Visit ESS-DIVE’s repository summary to see the full metric capabilities available to all portals

https://data.ess-dive.lbl.gov/profile
Creating Portals

- Any registered **Data Contributor** can create a portal!

- After you **login** to ESS-DIVE with your ORCiD and you can create a new portal from your settings page.
Creating Portals

- Any registered **Data Contributor** can create a portal!

- After you **login** to ESS-DIVE with your ORCiD and you can create a new portal from your settings page
Creating Portals

- **ESS-DIVE** has extensive **documentation** on how to develop your portal.

- Our DataONE partners at the Arctic Data Center have **tutorial videos** available.

Portals

A new feature on ESS-DIVE. Researchers can now easily view project information and datasets all in one place.

What is a Portal?

A portal is a collection of private or public ESS-DIVE data packages on a unique webpage.

Typically, a research project’s website won’t be maintained beyond the life of the project and all the information on the website that provides context for the data collection is lost. ESS-DIVE portals can provide a means to preserve information regarding the projects’ objectives, scopes, and organization and couple this with the data files so it’s clear how to use and interpret the data for years to come.

Portals also leverage ESS-DIVE’s metric features, which create statistics describing the project’s data packages. Information such as total size of data, proportion of data file types, and data collection periods are immediately available from the portal webpage.
Creating Portals

- ESS-DIVE has extensive documentation on how to develop your portal
- Our DataONE partners at the Arctic Data Center have tutorial videos available
Creating Portals

● ESS-DIVE has extensive documentation on how to develop your portal

● Our DataONE partners at the Arctic Data Center have tutorial videos available. Give it a try!
Are you or your group interested in creating a portal for your published data packages?

- Yes! [A] 100%
- Not at this time [B]
ESS-DIVE data and metadata reporting formats
Data stored in repositories should be FAIR

Findable
Accessible
Interoperable
Reusable

60% of publicly archived data is unusable.

(Wilkinson et al. 2016; Roche et al. 2015)
Clarifying terminology: Data standards and reporting formats

● **Data Standards** - Decades of development, international community of governance

● **Reporting Formats** - In development, though still enable data harmonization and synthesis
Non-standardized data are difficult to reuse and synthesize

<table>
<thead>
<tr>
<th>idNumber</th>
<th>material</th>
<th>temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>3928</td>
<td>soil</td>
<td>23.2</td>
</tr>
<tr>
<td>3234</td>
<td>groundwater</td>
<td>9.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>sampleNum</th>
<th>substance</th>
<th>temp</th>
</tr>
</thead>
<tbody>
<tr>
<td>8765</td>
<td>dirt</td>
<td>21.1</td>
</tr>
<tr>
<td>2312</td>
<td>ground liquid</td>
<td>7.0</td>
</tr>
</tbody>
</table>
Standardized data are more readily reused and synthesized

<table>
<thead>
<tr>
<th>idNumber</th>
<th>material</th>
<th>temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>3928</td>
<td>soil</td>
<td>23.2</td>
</tr>
<tr>
<td>3234</td>
<td>groundwater</td>
<td>9.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>idNumber</th>
<th>material</th>
<th>temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>8765</td>
<td>soil</td>
<td>21.1</td>
</tr>
<tr>
<td>2312</td>
<td>groundwater</td>
<td>7.0</td>
</tr>
</tbody>
</table>
Community-led reporting format development to make diverse ESS data FAIR

ESS-DIVE team developed high priority, base data formats

Developing reporting formats for select data types with $1M from DOE’s Biological Environmental Research (BER) program
ESS-DIVE’s GitHub Community Space is for reporting format, code, tools

ESS-DIVE Community Space

A workspace for code, tools, reporting formats and other products related to the ESS-DIVE repository. Please message ess-dive-support@lbl.gov to contribute

Berkeley, CA USA  http://ess-dive.lbl.gov  @ESSDIVE  ess-dive-support@lbl.gov

Repositories 6  Packages  People 4  Teams  Projects  Settings

Pinned repositories

essdive-community-space-guide
A guide to starting or contributing to ESS-DIVE community repositories

R
Joan Damerow (LBNL)
### ess-dive-community / essdive-sample-id-metadata

<table>
<thead>
<tr>
<th>Directory</th>
<th>File Details</th>
<th>Last Updated</th>
</tr>
</thead>
<tbody>
<tr>
<td>terms</td>
<td>Update sampleMetadata_sources.md</td>
<td>2 months ago</td>
</tr>
<tr>
<td>README.md</td>
<td>Update README.md</td>
<td>15 days ago</td>
</tr>
<tr>
<td>contribute.md</td>
<td>Update contribute.md</td>
<td>2 months ago</td>
</tr>
<tr>
<td>guide.md</td>
<td>Update guide.md</td>
<td>15 days ago</td>
</tr>
<tr>
<td>instructions.md</td>
<td>Update instructions.md</td>
<td>2 months ago</td>
</tr>
<tr>
<td>sampleTemplate.csv</td>
<td>Add files via upload</td>
<td>last month</td>
</tr>
<tr>
<td>sampleTemplate.xls</td>
<td>Add files via upload</td>
<td>last month</td>
</tr>
</tbody>
</table>

---

**Joan Damerow (LBNL)**
**Sample Description:**
Material* | Field name informal classification | Sample Description |

### Material

<table>
<thead>
<tr>
<th>Proposed ESS-DIVE Element</th>
<th>material</th>
<th>Required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Definition</td>
<td>Material that the sample consists of.</td>
<td></td>
</tr>
<tr>
<td>Format</td>
<td>SESAR controlled list. See ESS-DIVE’s proposed material terms from Environment Ontology (ENVO)</td>
<td></td>
</tr>
<tr>
<td>Additional Instructions</td>
<td>Use a semi-colon to delimit multiple materials where needed. ESS-DIVE is requesting additional terms for organisms, organic material, and water samples. Please provide feedback on any other terms needed.</td>
<td></td>
</tr>
<tr>
<td>Examples</td>
<td>soil; sediment; surface water ENVO:00002042; groundwater ENVO:01001004</td>
<td></td>
</tr>
</tbody>
</table>
GitHub “Issues” enable us to track suggestions to reporting format from user community

**Term to change**: Exact name of term you are suggesting change to.

**Submitter**: Provide your name.

**Justification**: Why is this change necessary?
Updates to GitHub instantly pushed to the GitBook

```markdown
# ESS-DIVE Sample ID and Metadata Reporting

ESS-DIVE recommends registering samples for [Global Sample Numbers](#)
```
Easy translation of GitHub content to user-friendly documentation

ESS-DIVE Sample ID and Metadata Reporting

ESS-DIVE recommends registering samples for Global Sample Numbers (IGSNs) through the System for Earth Sample Registration (SESAR). IGSNs are associated with standardized metadata to characterize a variety of different samples and their collection details. These sample identifiers facilitate sample discovery, tracking, and reuse; they are especially useful when sample data is shared with collaborators, sent to different labs or user facilities for analyses, or distributed in different data files, datasets, and/or publications.

https://ess-dive.gitbook.io/sample-id-and-metadata/
Learn more and contribute to reporting formats on
our GitHub Community Space

https://github.com/ess-dive-community

File-Level Metadata
CSV Files
Soil Respiration
Sample ID Metadata

Leaf Physiology
Hydrologic Monitoring
Water & Soil Chemistry
Amplicon sequencing
Questions?

Detailed guides are available at docs.ess-dive.lbl.gov

ESS-DIVE Community Space on GitHub https://github.com/ess-dive-community

Contact us at ess-dive-support@lbl.gov
ESS-DIVE Glossary

List of common ESS-DIVE terms:

- **DataONE** - The Data Observation Network for Earth (DataONE) is a distributed framework and sustainable cyberinfrastructure that provides open and secure access to Earth observational data. ESS-DIVE is a DataONE member.

- **DOE** - The U.S. Department of Energy (DOE) is a Cabinet-level department of the United States whose mission is to ensure America’s security and prosperity by addressing its energy, environmental and nuclear challenges through transformative science and technology solutions.
ESS-DIVE Glossary (cont.)

- **ESGF** - The Earth System Grid Federation (ESGF) is a collaboration that develops, deploys and maintains software infrastructure for the management, dissemination, and analysis of model output and observational data. It is an interagency and international effort led by the U.S. Department of Energy with various co-funding agencies.

- **ESS** - Environmental Systems Science (ESS) is a U.S. Department of Energy Office of Science program under the Biological and Environmental Research Program seeking to advance a robust predictive understanding of terrestrial surface and subsurface ecosystems.
ESS-DIVE - Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) is a U.S. Department of Energy repository for earth and environmental science data, models and software generated from research on terrestrial and subsurface environments.

NCEAS - The National Center for Ecological Analysis and Synthesis (NCEAS) at UC Santa Barbara partners with ESS-DIVE on data preservation. NCEAS is a DataONE member and a recognized expert in ecological data, digital libraries, and standards for data format.
ESS-DIVE Glossary (cont.)

- **NERSC** - The National Energy Research Supercomputing Center (NERSC) is the primary scientific computing facility for the Office of Science in the Department of Energy, and is the primary data storage center for ESS-DIVE.

- **ORCiD** - The Open Researcher and Contributor IDentifier provides anyone a persistent digital identifier (an ORCID iD) that distinguishes researchers from one another and provides a record that supports automatic links among all professional activities.