Using RESTClient for Firefox to test syntax and load data

The REST Client Add-On for Firefox is useful for testing the syntax of sbJSON packaged and presented to ScienceBase for loading and manipulating data. It provides a simple, visual way to be authenticated in ScienceBase. A user can view how the REST services will behave under different scenarios or perform actual data management tasks. These tasks are going to be used once so where building a full application of some kind is not really necessary. The following example describes a real case where information was packaged and loaded into ScienceBase to create new items

Use Case: One-time load of item data to ScienceBase

The following example came from work to address a challenge presented by the research director of the National Research Program in the USGS Water Mission Area. The NRP currently has a web site containing the research record of projects and publications produced by program scientists since about 1990. The site contains listings of citations (some with abstracts and/or links), project descriptions and their associated citations, listings of research scientists in the program, and an organization that includes grouping by research focus areas. The challenge presented to the Community for Data Integration was to get the information into a database where it could be more readily maintained and have more flexibility in how it could be accessed and used. ScienceBase provides part of this solution using its extensions for citations and projects along with the inherent information in the ScienceBase Directory to house and present the information and relationships between researchers, projects, and products. Information on projects needed to be stored natively in ScienceBase, while an appropriate citation management solution provides the best option for managing and organizing raw citations, which can later be loaded and synchronized with ScienceBase to complete the "database or NRP projects, citations, and researchers."

Data preparation

In this scenario, it was desirable to capture project titles and descriptions, principal investigators, and "Areas of Investigation: Current Research Topics" and insert the project records, complete with those details, into ScienceBase. The current topics page (http://water.usgs.gov/nrp/currenttopics.html) was used to start a screen scraping run to grab up the basic information on each project (title and abstract listed at the top of each project page) along with principal investigators (listed at the bottom of most project pages) and the list of terms used to describe research areas. Any number of different applications or scripting libraries could be used for this exercise. The end result here was some cleaned up text in a spreadsheet of projects, including the following attributes:

- Title
- Abstract (body)
- URL of the web page where the information was extracted ("Original Web Page")
- Terms in a list (areas of investigation)
- Principal Investigators (including name, email, and phone number)

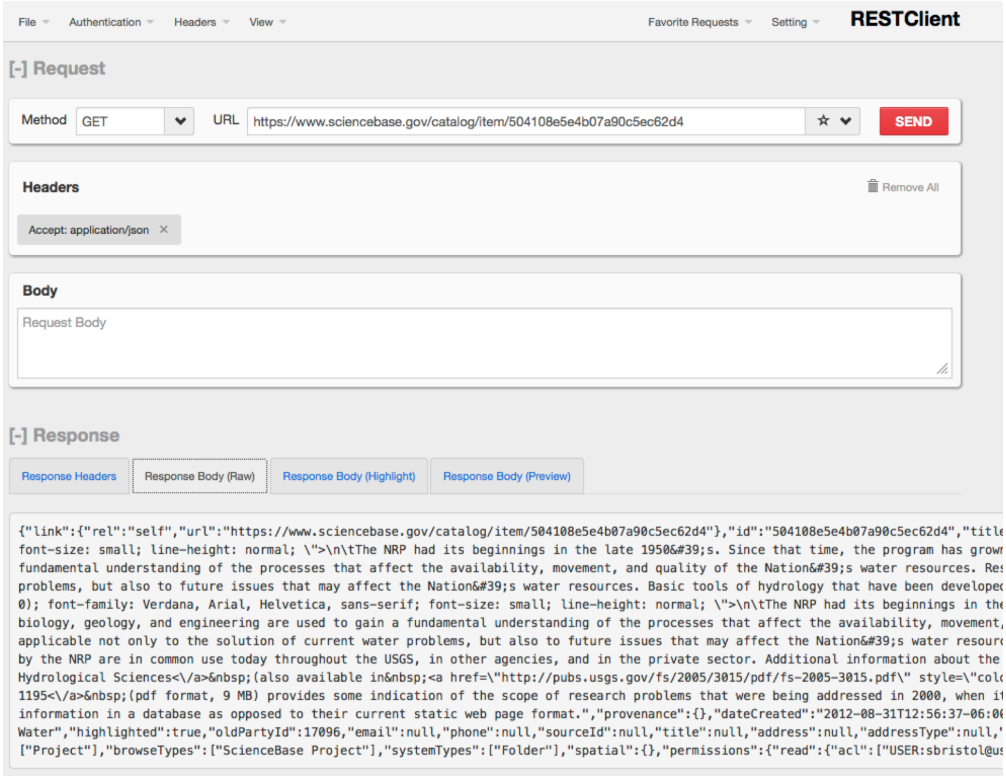
A little bit of additional work was done to lookup the IDs of all of the NRP researchers from the references so those could be directly linked to ScienceBase Directory entities instead of relying on the matching algorithm that is still in development. We also made a connection to the Office of the Chief Scientist for Water for all of these records as an organizational contact, but a more direct connection could be made from these records to one of the three regional research branches.

GET an item and the sbJSON syntax

All of the interactions with the REST services that involve more than simply returning an item or search result will involve presenting ScienceBase with JSON in the sbJSON syntax discussed elsewhere in documentation. Before working with an example like the above, it is valuable to look over some examples of existing records in sbJSON using the GET method. Simply browsing ScienceBase for similar items and using the "View as JSON" feature will show the raw JSON output, or you can use the RESTClient add-on as a good way to get started with that tool.

> To work with the sbJSON format in the ScienceBase REST service, you will either need to add an "Accept" header to the requests or add the "format=json" to the URL string. It's easy enough to add the Accept header with a value of "application/json" to the Firefox RESTClient, so that is the recommended method in this case.

Set the Method to Get, paste the URL to an item of interest into the URL field, and push the Send button in the RESTClient, and you will see the Response below with several tabbed elements and view options.



The sbJSON syntax is fairly straightforward, and one of the nice things about JSON is that it is generally more human readable than comparable XML formats, the raw response body seen in the above example notwithstanding. You will find much information on the web about JSON, including the http://www.json.org/ site. In working with a use case such as described here, it was useful to use a JSON validator to both check proper formatting on the syntax and to create a more readable layout of the text. There are many validators available, including those built into popular tools, but there are freely available online versions that work quite well such as JSONLint.

Creating a new item from RESTClient

Authentication

You must be logged in to create or edit items. You will get a Status Code of 200 in the RESTClient even if you are not logged in, but the item WILL NOT be created and it will return Formatted HTML (in the form of a login page) instead of the Formatted JSON of your new item. The RESTClient add-on allows you to sign into USGS in a different tab.

To create a new item using the RESTClient for Firefox we first set the "Method" to "POST" and the URL to point at "www.sciencebase.gov/catalog/item/". The header will be set to "Content-Type". The value must now be set to "application/json" since we are sending JSON. We put our JSON package into the "Body." Very minimal information is needed to create a basic ScienceBase item - only a parentId, for which your user credentials have "write" access, and a title.

```
{"parentId":"4f28888de4b050c1ade5f38e","title":"Test Item Title"}
```

Creating a more complete item

Refer to the documentation on ScienceBase items and extensions to see the full sbJSON syntax. You may also want to output a full example from an existing item in ScienceBase to see how specific types of values are encoded. The following is a full example of an NRP project record created in this use case.

Full Example

```
{
    "parentId": "504108e5e4b07a90c5ec62d4",
    "title": "Field Applications of Unsaturated Zone Flow Theory",
    "body": "Various processes within the unsaturated zone affect ground-water availability and portability, as
```

well as concentrations of water vapor and trace gases in the atmosphere. The rate at which precipitation or applied irrigation water infiltrates, its redistribution following infiltration, and the partitioning of the redistributed soil moisture between ground-water recharge and evapotranspiration affect the rate at which the ground-water reservoir is replenished and the degree to which ground water might be contaminated by chemical applications, spills, or disposal. Consequently, knowledge of and methods to quantitatively measure and predict these processes are needed to determine the impact of such societal practices as irrigation development for agriculture, the use of agricultural chemicals, and the disposal of radioactive and/or hazardous waste in the unsaturated zone on both the availability and potability of ground water. Processes governing transport in the unsaturated zone gas phase are also important in determining the potential for ground- water contamination by volatile compounds, the rate at which water is returned from soil moisture to the atmosphere as vapor, and the fate of other 'greenhouse gases', such as carbon dioxide, methane, and chlorofluorocarbons (CFCs). An understanding and quantification of these processes is needed both to assess the hazards of ground-water pollution and to better predict the impact of global change on future climate. Also see Unsaturated Zone Hydrology.",
    "contacts": [
        {
            "oldPartyId": 17096,
            "type": "Contact",
            "name": "Office of the Chief Scientist for Water",
            "highlighted": "true"
        },
        {
            "oldPartyId": 11034,
            "type": "Principal Investigator",
            "name": "Edwin P. Weeks",
            "email": "epweeks@usgs.gov",
            "phone": "303-236-4981"
        },
        {
            "oldPartyId": 4383,
            "type": "Principal Investigator",
            "name": "Richard W. Healy",
            "email": "rwhealy@usgs.gov",
            "phone": "303-236-5392"
        },
        {
            "oldPartyId": 209,
            "type": "Principal Investigator",
            "name": "Dean E. Anderson",
            "email": "deander@usgs.gov",
            "phone": "303-236-5691"
        },
        {
            "oldPartyId": 10422,
            "type": "Principal Investigator",
            "name": "James A. Tindall",
            "email": "jtindall@usgs.gov",
            "phone": "303-236-5005"
        },
        {
            "oldPartyId": 9835,
            "type": "Principal Investigator",
            "name": "David I. Stannard",
            "email": "distanna@usgs.gov",
            "phone": "303-236-4983"
        }
    ],
    "provenance": {
        "html": "Record was created from original web page content (see link)."
    },
    "webLinks": [
        {
            "type": "webLink",
            "typeLabel": "Web Link",
            "title": "Original Web Page",
            "uri": "http://water.usgs.gov/nrp/proj.bib/weeks.html"
        }
    ],
    "browseCategories": [
        {
            "name": "Project"

```
        }
    ],
    "browseTypes": [
        {
            "name": "ScienceBase Project"
        }
    ],
    "tags": [
        {
            "name": "Arid Land Hydrologic Studies"
        },
        {
            "name": "Carbon Cycle"
        },
        {
            "name": "Contaminant Reactions and Transport"
        },
        {
            "name": "Evapotranspiration"
        },
        {
            "name": "Ground-Water Flow, Transport, and Reactions"
        },
        {
            "name": "Hydroclimate (including droughts and floods)"
        },
        {
            "name": "Hydrogeology"
        },
        {
            "name": "Unsaturated Zone"
        },
        {
            "name": "Volcanic Hazards"
        }
    ]
}
```

Creating multiple items

Adding more than one item with one REST post is as simple as including multiple JSON items packaged as above but in a JSON array. Multiple items also need to be posted to the /items/ controller instead of the /item/ controller, meaning that the URL in the RESTClient needs to be set to https://www.sciencebase.gov/catalog/items/. A simple example of the sbJSON syntax for multiple items is included below:

```
[
    {
        "parentId": "504108e5e4b07a90c5ec62d4",
        "title": "Title of item 1"
    },
    {
        "parentId": "504108e5e4b07a90c5ec62d4",
        "title": "Title of item 2"
    }
]
```

In the NRP project case, we added sbJSON syntactical elements to a concatenation function in the spreadsheet of values scraped from the web pages to build a JSON package of 59 new items. These were posted with one call to the REST service using the RESTClient add-on, placing the items into a prepared parent item (folder) to contain the projects and provide context. The entire process took less than 2 seconds to insert what came to 2,972 lines of "prettied" JSON.

The ScienceBase REST service for inserting multiple items has been load tested with several hundred items at a time with reasonable results, but most working clients that are harvesting records into ScienceBase are executing multiple consecutive post processes. ScienceBase write access is restricted to a relatively small group of known users building client applications. Please do be kind to the service and check in at sciencebase@usgs.gov if you are planning a large operation of some kind. It is also a good idea to test new operations in beta before attempting on the ScienceBase production system.

Updating items

Updating items in ScienceBase is handled in similar fashion in terms of the sbJSON syntax passed to the service. You can update all or part of a given item by using the PUT HTTP method and sending the request to the item ID you want to update. If you send an update for a particular part of an item that can contain multiple values (e.g., tags used to describe the item), you will need to send the full package - meaning that you will need to include all of the values you want for the final element. For instance, if you want to add a tag to an existing set of tags, you will need to include the current tags and the new tag in the JSON package sent to an item.

In our use case, we wanted to add all of the "USGS NRP Research Topics" used in all of the 59 project records to the parent folder containing all of the projects so that we could have a nice clickable list of topics to browse the projects. We used the following sbJSON syntax, PUT to the folder item:

```
{
    "tags": [
        {
            "name": "Acid Mine Drainage",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Aquatic Habitat",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Arid Land Hydrology",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Carbon Cycle",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Contaminant Reactions and Transport",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Estuaries and Near-Shore Environments",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Evapotranspiration",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Fire effects on watersheds",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Fractured Rock",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Ground-Water Flow, Transport, and Reactions",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Ground-water Hydraulics",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Hydroclimate (including droughts and floods)",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Hydrogeology",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Hypoxia (see nutrients)",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Invasive Species",
```

```
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Isotopic Tracers",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Lakes",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Metals",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Microbiology: Activity, Biogeochemistry, and Transport in Water",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Nutrients",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Organic Compounds (natural and man-made)",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Radionuclides",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Rivers and Streams",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Sediment Chemistry",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Sediment Transport",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Statistical Hydrology",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Stream Channel Morphology",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Surface Chemistry",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Surface-Water - Ground-Water Interactions",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Surface-Water Hydraulics",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Surface-Water Transport and Reactions",
        "type": "USGS NRP Research Topics"
    },
    {
        "name": "Unsaturated Zone",
        "type": "USGS NRP Research Topics"
    },
    {
```

```
            "name": "Volcanic Hazards",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Watershed Studies",
            "type": "USGS NRP Research Topics"
        },
        {
            "name": "Wetlands",
            "type": "USGS NRP Research Topics"
        }
    ]
}
```


Tags Schema

The schema for "Tags" in ScienceBase is currently evolving, but it includes quite a number of additional elements than those demonstrated in the above example. We are working on developing a more robust vocabulary system that allows for both formal, controlled vocabularies and community- or collection-specific vocabularies, complete with all of the elements needed for both SKOS and OWL representations. The example from this use case simply added a type to these terms to help distinguish them from similar terms coming from other collections of records. In the iteration of ScienceBase at the time of writing, all topical tags from any named vocabulary are simply put together under the broad notion of "topics" as we work through the architecture of how to better handle vocabularies across the diversity of items in ScienceBase.

Extending into more complex uses

This documentation glosses over a number of issues that we worked through in formatting the information about the NRP projects and validating that we had correct details that would be accurately displayed in ScienceBase. The majority of work in a case like this will be in preparing the data and ensuring that the correct sbJSON syntax is applied for all of the elements of the ScienceBase Item Model and any extensions that need to be populated. However, once those details are ironed out, the actual process of submitting a large batch of records or changes is quite straightforward as described above.

In this case, we chose to work the information from the standpoint of a relatively non-sophisticated data handler, using a simple spreadsheet (Excel) to organize and prepare the sbJSON package. The idea was to use tools and methods that should be well within the grasp of a desktop software savvy user who might be able to work through the same overall process for similar use cases where a batch of records needs to be loaded. We first wrote a Python script using BeautifulSoup for the web scraping process but also ran it through using a graphical tool called Outwit. Both methods worked well, but the Outwit tool made the process quite simple for a more visual information manager. The greatest challenge in any web scraping (unstructured information to structured data) application like this is dealing with the wide variability in formatting that is often in place with static web sites. The next steps in this particular use case are more challenging - converting 6,000+ citation records, at varying levels of completeness and format, into structured citation data for import and management in a citation management system such as EndNote (and eventual synchronization with ScienceBase using a more complex process).

ScienceBase REST services make for a very flexible system in terms of methods that can be applied to manage information in ScienceBase. Extending beyond the few core attributes handled in this example, REST services can be used to manage any of the ScienceBase extensions and to handle file uploads and manipulation of files in the ScienceBase Repository.