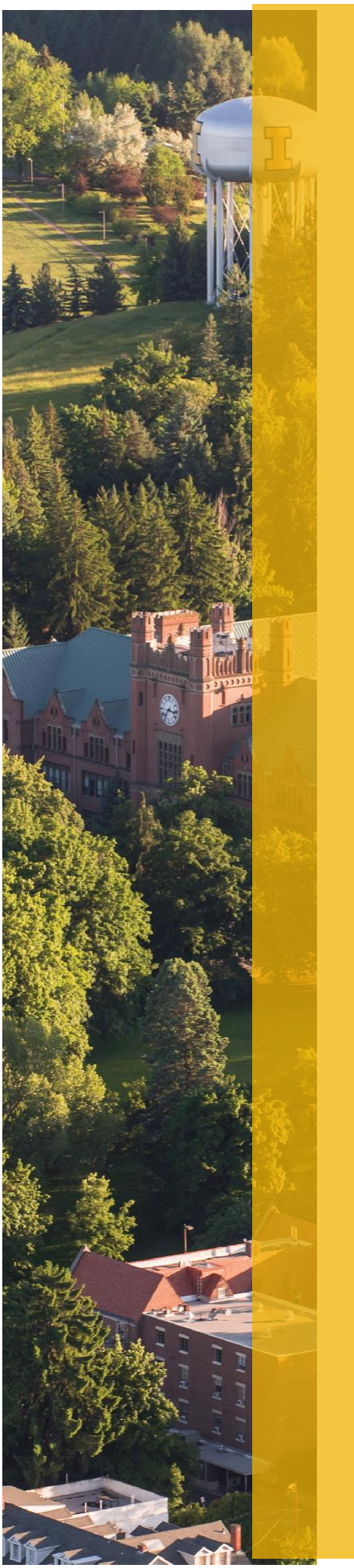
# Assignment 2

## Data Report

2022

University of Idaho

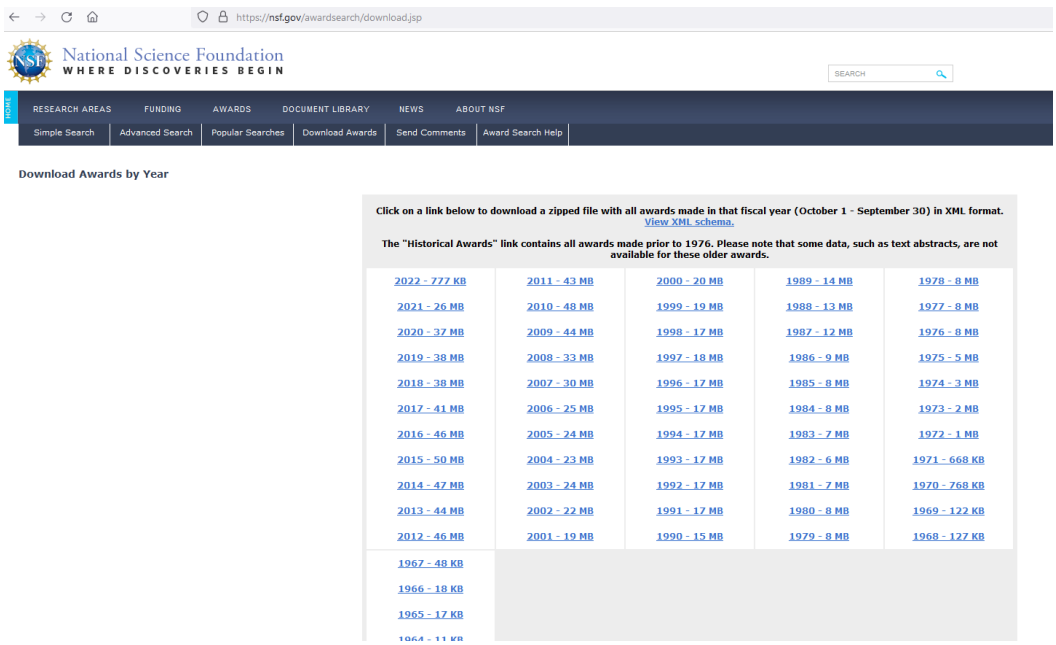




# GOAL:

- The goal or objectives of the Data collection of Awards which is an open data from National Science Foundation (NSF) are -
  - to evaluate the number of actual awards granted by NSF.
  - to capture the quality evidence of the amount of these awards on different researches.
  - to investigate and compare the different departments in which researches are going on.
  - to observe the trend and pattern of awards over the year.
  - to analyse the relationship among the above stated factors and trace out some insights.

- The data is about research projects that NSF has funded since 1989 which is in XML format and is collected from –




- Schema of the dataset is-

```
<?xml version="1.0"?>
<xsd:schema attributeFormDefault="unqualified" elementFormDefault="qualified" version="1.0" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <xsd:element name="rootTag">
    <xsd:complexType>
      <xsd:sequence>

        <xsd:element name="Award">
          <xsd:complexType>
            <xsd:sequence>
              <xsd:element name="AwardTitle" type="xsd:string" />
              <xsd:element name="AwardEffectiveDate" type="xsd:dateTime" />
              <xsd:element name="AwardExpirationDate" type="xsd:dateTime" />
              <xsd:element name="AwardAmount" type="xsd:int" />

              <xsd:element name="AwardInstrument">
                <xsd:complexType>
                  <xsd:sequence>
                    <xsd:element name="Value" type="xsd:string" />
                  </xsd:sequence>
                </xsd:complexType>
              </xsd:element>

              <xsd:element name="Organization">
                <xsd:complexType>
                  <xsd:sequence>
                    <xsd:element name="Code" type="xsd:int" />
                    <xsd:element name="Directorate">
                      <xsd:complexType>
                        <xsd:sequence>
                          <xsd:element name="LongName" type="xsd:string" />
                        </xsd:sequence>
                      </xsd:complexType>
                    </xsd:element>
                    <xsd:element name="Division">
                      <xsd:complexType>
                        <xsd:sequence>
                          <xsd:element name="LongName" type="xsd:string" />
                        </xsd:sequence>
                      </xsd:complexType>
                    </xsd:element>
                  </xsd:sequence>
                </xsd:complexType>
              </xsd:element>

              <xsd:element name="ProgramOfficer">
                <xsd:complexType>
                  <xsd:sequence>
                    <xsd:element name="SignBlockName" type="xsd:string" />
                  </xsd:sequence>
                </xsd:complexType>
              </xsd:element>

              <xsd:element name="AbstractNarration" type="xsd:string" />
```

## Mode of collection:

- The mode of data collection is secondary data, and the mode is driven by exploration and reporting of the existing source data via downloading existing data from the above-mentioned portal.

- The need is driven by exploration of the information government seek regarding the research NSF funds in science & engineering. The same has been provided to the public through the Public Access Policy.

- These types of data collection modes often lead to new investigative questions which further helps other organizations for analysis with the sampling of data.

## Actual Data Derivation:

- The actual data is derived by downloading the zip files from the NSF's award site.

- There were no challenges as such. The data will be directly parsed from XML to normal human readable format.

- There are multiple values for a single field which is an issue to store. It could store in an array like below –

```
<ProgramReference>
<Code>7363</Code>
<Text>RES IN NETWORKING TECH &amp; SYS</Text>
</ProgramReference>
<ProgramReference>
<Code>8228</Code>
<Text>CISE Resrch Initiatn Initiatve</Text>
</ProgramReference>
```

Data for a single field is represented by two values viz 7363 & 8228.

- Also, few of the fields are in CDATA format –

```
<Code>01001920DB</Code>
<Name><![CDATA[NSF RESEARCH & RELATED ACTIVIT]]></Name>
<FUND_SYMB_ID>040100</FUND_SYMB_ID>
```

- Further, the data is separated by financial year directory. In turn, processing that data is tedious and manual task which has a high future scope.

## Physical & Logical Organization:

- **Logical Organization** of data –

- The data is of Awards granted, therefore, creating an ontology will help us in future perspective too.

- The ontology will be considering each element as object and how fields are related to each other.
- **Object Data ontology will organize the following base formats and structures** for each data field values
  - i. Entity information – entity_
  - ii. Target Entity in relationship – target_entity_
  - iii. Relationship information – relationship_
  - iv. Time – time_
    - 1. time_document will serve as the master time field which means the time of the dataset.
    - 2. There could be several metadata time fields like time_collected, time_uploaded, and time_updated.
  - v. Metadata information – doc_

- This ontology will help us in index naming strategies and hence, make it easier to identify the type of data in future.
- The attached below mapping enlists the source name to the corresponding ontological name.

mapping_nsf.xlsx

- **Physical Organization** of data –

  - The data is downloaded from the site in XML format and kept as raw in my GitHub directory.

  - The data will get processed by the script (written in logstash) and the desired data will get stored in the ElasticSearch Index (main_nsf).

```
GET main_nsf/_search
```

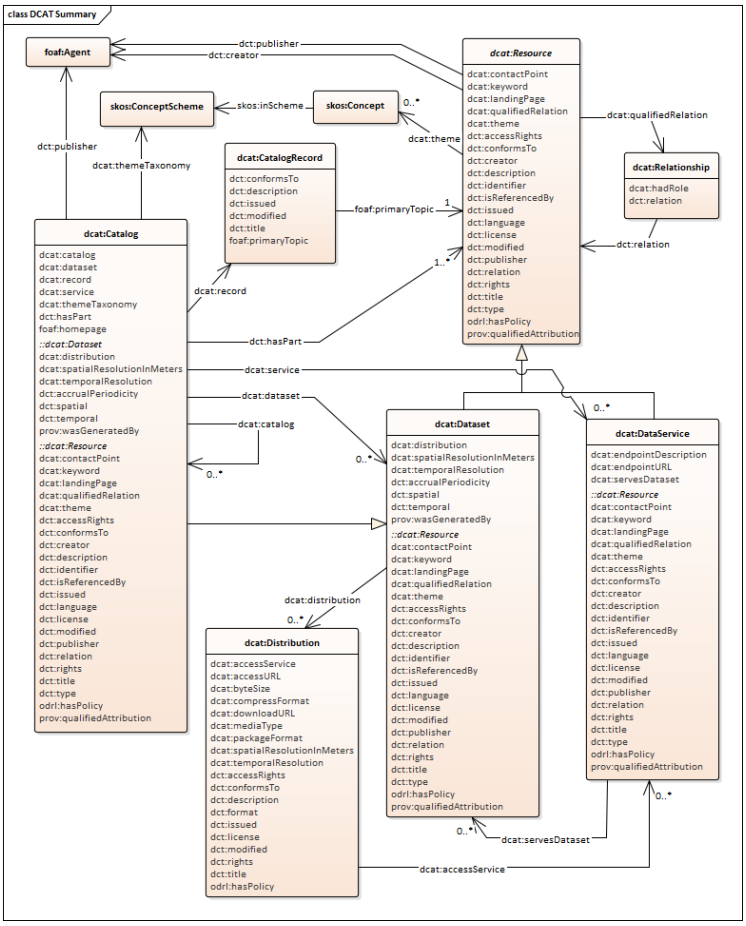

```
16          "_type" : "_doc",
17          "_id" : "08090000-2021-10-01T00:00:00.000Z-2200033-M
              Laboratory-50665-2022-02-13T07:38:42.463Z",
18          "_score" : 1.0,
19          "_source" : {
20            "message_current_total_award_value" : "50665",
21            "entity_parent_award_id_piid" : "2200033",
22            "contract_research_type" : "Animal Developmental M
23            "entity_awarding_office_code" : "BIO",
24            "entity_primary_pop_name" : "Marine Biological Lab
25            "entity_alias" : "MARINE BIOLOGICAL LABORATORY, TH
26            "entity_primary_pop_zip_code" : "025431015",
27            "relationship_type" : "grant from",
28            "entity_primary_pop_congressional_district" : "09"
29            "entity_awarding_office_name" : "Direct For Biolog
30            "entity_phone" : "15082897243",
31            "entity_country" : "United States",
32            "entity_state_code" : "MA",
33            "entity_duns" : "001933779",
34            "entity_primary_pop_address" : "7 MBL Street",
35            "message_title" : "EDGE: Generating Tools to Study
                Development",
36            "doc_cfda_id" : "47.074",
37            "relationship_details" : "Standard Grant",
38            "entity_zip_code" : "025431015",
39            "entity_congressional_district" : "MA09",
40            "time_uploaded" : "2022-02-13T07:38:42.463Z",
41            "entity_funding_agency_code" : "4900",
42            "entity_awarding_sub_agency_code" : "08090000",
43            "time_filing_period_start" : "2021-12-06T00:00:00.
44            "entity_officer" : {
```


- The schema mapping as given below is required while storing the data in the datalake of ElasticSearch.


```
{
  "main_nsf" : {
    "aliases" : { },
    "mappings" : {
      "_doc" : {
        "dynamic" : "true",
        "properties" : {
          "contract_action_type" : {
            "type" : "keyword"
          },
          "contract_actions_value" : {
            "type" : "long"
          },
          "contract_award_a76_fair_act_action" : {
            "type" : "keyword"
          },
          "contract_award_arra_value" : {
            "type" : "double"
          },
          "contract_award_clinger_cohen_act_planning" : {
            "type" : "keyword"
          },
          "contract_award_consolidated" : {
            "type" : "keyword"
          },
          "contract_award_construction_wage_rate_requirements" : {
            "type" : "keyword"
          },
          "contract_award_contingency_humanitarian_or_peacekeeping_operation" : {
            "type" : "keyword"
          },
```


- The Foundation maintains a public data listing, where each data set is described using a metadata profile that corresponds to the Data.gov common core standard.

- The metadata schema and Data Standards **are** specified in for the dataset is based on DCAT, a hierarchical vocabulary specific to dataset




- The Data Standards are derived from the international W3C DCAT-US.

## Metadata & Documentation :

- Metadata is structured information of data enlisted that describes, explains, locates, and makes it easier to retrieve, use, or manage an information resource.

- Provenance is identified as a derivative of Data.gov common core standard which is in turn is based on the international W3C DCAT-US specification.

- The documentation is provided at https://resources.data.gov/resources/dcat-us/

- The quality standard of the data is maintained and elaborated at https://www.nsf.gov/policies/infoqual.jsp

- The link for the metadata enlisted will be https://resources.data.gov/resources/podm-field-mapping/#field-mappings.

ISO 19115 to Project Open Data Metadata Schema v1.1 (POD v1.1) Crosswalk

| POD v1.1 | Label | Condition | Repeats | Guidance | ISO Description | ISO XPath |
|---|---|---|---|---|---|---|
| **Catalog Fields** | | | | | | |
| conformsTo | Schema Version | required | | populated by Agency Enterprise Inventory Application | – | – |
| dataset | Dataset | required | | populated by Agency Enterprise Inventory Application | – | – |
| **Dataset Fields** | | | | | | |
| title | Title | required | no | | title | `//gmd:identificationInfo /gmd:MD_DataIdentification /gmd:citation /gmd:CI_Citation /gmd:title /gco:CharacterString` |
| description | Description | required | no | | abstract | `//gmd:identificationInfo /gmd:MD_DataIdentification /gmd:abstract /gco:CharacterString` |

- Under the OPEN Government Data Act and the Open Data Policy, federal agencies are required to publish an enterprise data inventory, provided as a JSON file, using the standard DCAT-US metadata schema and hosted on an agency's website at agency.gov/data.json.

## Data Management Plan Experience :

- **Logical Collections:**
    - This is ontology I have defined as per my understanding of the data.
    - The object ontology worked just fine as I would be able to draw relationships and per the future scope as well.

- **Physical Data:**
    - Replication, Backing Up & Caching: The raw data is an open dataset and hence, no replication is as such required. Since the data flow is not dynamic or streaming, there is no need for caching or backing up. I could run my data processing script anytime and get the o/p but I have

stored the visualized data in my local Kibana System to bring in on some insights.

- Physical Storage: The data will be uploaded to my GitHub repository. The data is in XML format loaded in ElasticSearch(in Json format) datalake
- Security: Since the data and the visualisations made will be uploaded in my GitHub repository, it is password protected and further only the authorized person will be able to access in future even though the data is right now is publicly available.
- Data Format: The data is in XML (semi-structured) format and will be processed and converted into Json

- **Interoperability support:**
  - As raw and reporting data will be made public along with the processing script, it will support interoperability. Other agencies, providers, individuals etc., can use as their part of research or understand different dimensions.
  - But the data is available on Data.gov and defined as per the common core standard which is in turn is based on the international W3C DCAT-US specification.
  - Only the required access needs to be given, and the data exchange can be at all stages in real-time.

- **Security Support:**
  - This was also not that challenging as the data access in the GitHub repository can be changed anytime.
  - But yes, this won't be real time available as the processing is being done in my system and direct the final o/p will be stored in the repository (only the data vizualisation will be available).

- **Data ownership:**
  - The primary owner of this data is National Science Foundation, a federal agency of U.S. Responsibilities of any changes in the data will be solely on NSF.

- **Metadata collection, management and access:**
  - The documentation is provided at https://resources.data.gov/resources/dcat-us/
  - The quality standard of the data is maintained and elaborated at https://www.nsf.gov/policies/infoqual.jsp
  - The metadata collections are not an issue as these are standardized already.

- **Persistence:**
  - The open data would reside in the GitHub repository for now. And then, it will get loaded in my local ElasticSearch as a database. After that, the final processed data will be either in Kibana after taking the visualizations for public use.

- **Discovery:**
    - Discovery at any stage can be done by public listed data or by my GitHub repository. Hence, no maintenance is required. Thus, no challenges are there.

- **Data dissemination and publication:**
    - Data dissemination and publication is done by data.gov.
    - Also, since the data will be publicly available, any one can add or change the data so that the same can be seen in the final o/p. The provisions will be made to incorporate the changes.
    - But, the data publication is still quite challenging and did not go well.