

Assignment 1

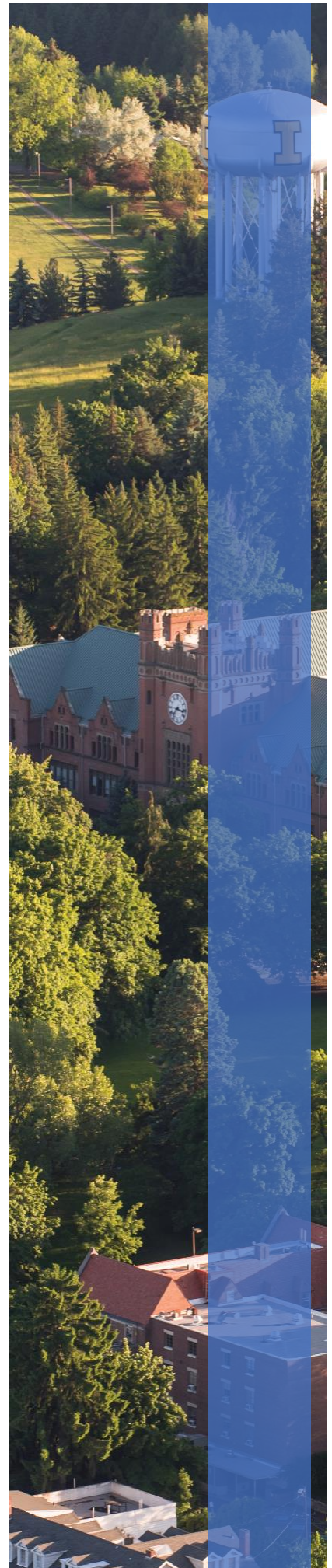
Data collection proposal

2022



University
of Idaho

It is U of I policy to prohibit and eliminate discrimination on the basis of race, color, national origin, religion, sex, sexual orientation and gender identity/expression, age, disability, or status as a Vietnam-era veteran. This policy applies to all programs, services, and facilities, and includes, but is not limited to, applications, admissions, access to programs and services, and employment.




Data Collection

Proposal:

1. Data are representation of facts and to get meaningful insights out of that is information. But to discovering data is itself a first step in data management.
2. As a part of this assignment, I am collecting the open data from [NSF](#) portal for timeframe 2021 & 2022 as of now. The data is about [Awards](#) as research projects that NSF has funded since 1989.

← → ↻ 🏠 nsf.gov/awardsearch/download.jsp 🔍

 **National Science Foundation**
WHERE DISCOVERIES BEGIN

RESEARCH AREAS FUNDING AWARDS DOCUMENT LIBRARY NEWS ABOUT NSF

Simple Search Advanced Search Popular Searches Download Awards Send Comments Award Search Help

Download Awards by Year

Click on a link below to download a zipped file with all awards made in that fiscal year (October 1 - September 30) in XML format. [View XML schema.](#)

The "Historical Awards" link contains all awards made prior to 1976. Please note that some data, such as text abstracts, are not available for these older awards.

2022 - 594 KB	2011 - 43 MB	2000 - 20 MB	1989 - 14 MB
2021 - 25 MB	2010 - 48 MB	1999 - 19 MB	1988 - 13 MB
2020 - 37 MB	2009 - 44 MB	1998 - 17 MB	1987 - 12 MB
2019 - 38 MB	2008 - 33 MB	1997 - 18 MB	1986 - 9 MB
2018 - 38 MB	2007 - 30 MB	1996 - 17 MB	1985 - 8 MB
2017 - 41 MB	2006 - 25 MB	1995 - 17 MB	1984 - 8 MB
2016 - 46 MB	2005 - 24 MB	1994 - 17 MB	1983 - 7 MB
2015 - 50 MB	2004 - 23 MB	1993 - 17 MB	1982 - 6 MB
2014 - 47 MB	2003 - 24 MB	1992 - 17 MB	1981 - 7 MB
2013 - 44 MB	2002 - 22 MB	1991 - 17 MB	1980 - 8 MB
2012 - 46 MB	2001 - 19 MB	1990 - 15 MB	1979 - 8 MB
1978 - 8 MB	1967 - 48 KB		

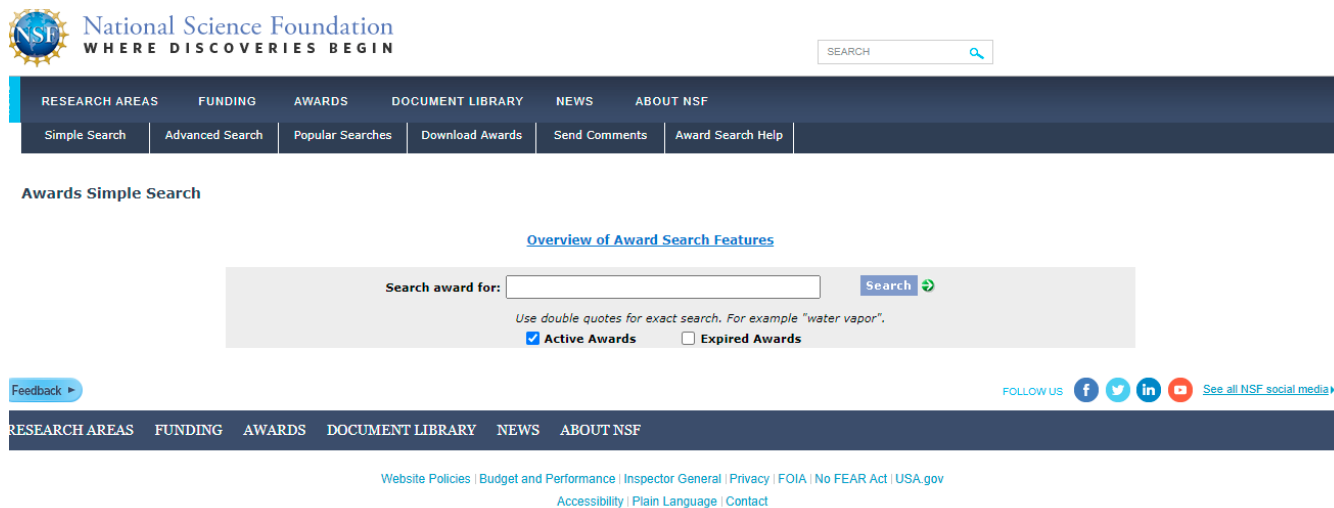
THE xml award dataset is type –

```
<?xml version="1.0" encoding="UTF-8" ?>
<rootTag>
  <Award>
    <AwardTitle>Computable Mathematics</AwardTitle>
    <AGENCY>NSF</AGENCY>
    <AwardEffectiveDate>06/01/2005</AwardEffectiveDate>
    <AwardExpirationDate>05/31/2008</AwardExpirationDate>
    <AwardTotalIntnAmount>105000.00</AwardTotalIntnAmount>
    <AwardAmount>105000</AwardAmount>
    <AwardInstrument>
      <Value>Standard Grant</Value>
    </AwardInstrument>
    <Organization>
      <Code>03040200</Code>
    </Organization>
    <Directorate>
      <Abbreviation>MPS</Abbreviation>
      <LongName>Direct For Mathematical & Physical Scien</LongName>
    </Directorate>
    <Division>
      <Abbreviation>DMS</Abbreviation>
      <LongName>Division Of Mathematical Sciences</LongName>
    </Division>
    <ProgramOfficer>
      <SignBlockName>Tomek Bartoszynski</SignBlockName>
      <PO_EMAIL>tbartosz@nsf.gov</PO_EMAIL>
      <PO_PHON>7032924885</PO_PHON>
    </ProgramOfficer>
    <AbstractNarration>In this project the principal investigator will apply methods from<br/>computability theory and other branches of logic and theoretical computer<br/>science to study the effective content and proof-theoretic
```

3. The type of data collection is secondary data, and the mode is driven by exploration via

downloading existing data from online portal. The main purpose is to determine how many and in which areas the major research funding has been done and understand the data timeline.

4. Data collection need is driven by exploration of the information government seek regarding the research NSF funds in science and engineering. The same has been provided to public through Public Access policy.
5. It includes abstracts that describe the research, and names of principal investigators and their institutions. The data repository includes both completed and in-process research.



Data Management:

1. Logical Collections:

- a. Since the data is about grants/awards, we can create an object ontology considering all elements as objects related to each other.
- b. Object Data ontology is described 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_

2. Physical Data: The actual data is downloaded from NSF's award site.

- a. 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. We can run our data processing script anytime and get the o/p.
- b. Physical Storage: The data will be uploaded in my GitHub repository. The data is in XML format and can be manipulated as per the desired o/p to get information.
- c. Security: Since the data and the visualisations made will be uploaded in my GitHub repository, it is password protected and further only the authorised person will get the access when required. Once the raw data finally gets converted for the reporting purposes, it can be made public to benefit others.
- d. Data Format: The data is in XML (semi-structured) format and will be processed by either

Python or Logstash.

- e. Naming Conventions- Since the data will follow the object ontology, they will be related to each other.

3. Interoperability support:

- a. As raw and reporting data will be made public, it will support interoperability. Other agencies, providers, individuals etc., can use as their part of research or to understand different dimensions.
- b. Only the required access needs to be given, and the data exchange can be at all stages at real time. Any changes or addition of data will be supported by the data pipeline as the data will be monitored on the daily basis. Hence, data will be in congruent with the ones given at website.

4. Security support:

- a. Except for data pipeline, everything will be made publicly accessible since raw data is open data set only. The final o/p having all the insights will also be public as to help society in their research work.
- b. However, the data pipeline will not be accessible except the required person as repository access can be given. As the pipeline for processing of data is not important at user end, hence public access is not required.
- c. The final o/p can be made available in my Tableau profile, hence there would not be any issue regarding the access or loss of the data. As there are chances of system crash, it is not advisable to keep data in local system.

5. Data ownership:

- a. 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.
- b. But the reporting data after manipulation is done will be owned by me as I have derived the o/p from raw data.

6. Metadata collection, management and access:

- a. TITLE: Awards
- b. CREATOR: National Science Foundation
- c. DESCRIPTION: The awards are research projects that NSF has funded which describe names of principal investigators and their institutions. The database includes both completed and in-process research.
- d. FORMAT: The data is represented in unstructured format. New meta elements could be add in future.
- e. IDENTIFIER: (URL) <https://nsf.gov/awardsearch/download.jsp>
- f. LANGUAGE: English
- g. DATE: 01/28/2022 0:55:04

7. Persistence:

- a. The data would reside in the GitHub repository for now. And then will get loaded in my local ElasticSearch as a database. After that, the final processed data will be either in Tableau or Kibana for public use.
- b. As data keeps on changing, the final o/p will get change accordingly. The persisted data will be in the data format.
- c. The data will be fetched and stored as per the time line of yearly frequency.

8. Discovery:

- a. People can discover the data as it an open dataset and the GitHub repository will be public then. Also, after the processing will get complete the same will be publicly accessible (made available in Tableau public profile or Kibana via some reports or visualizations).

9. Data dissemination and publication:

- a. The data will get updated as I will monitor on daily basis. Further, my pipeline will run daily basis so to capture the latest change. It will be in sync with the latest data.
- b. 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.

Survey of data storage/ formats

- a. Existing suitable formats are HTML, CSV, PDF etc. In NSF's website we can easily find these options as data formats for Grants, Publications etc. And all these are easy to store.
- b. The above-mentioned formats can be easily stored and parsed by python, ruby etc.
- c. The data storage or maintenance of such formats are easy, and they are of ASCII or ISO. Hence these data formats can be used global wide by public.

Survey of metadata conventions, standards

- a. The standard convention is GILS focuses on Government/organizations. The metadata is owned and maintained by the owner of the award institution.
- b. But there is no proper format for metadata. However, the field information is stored as a part of Data. In the NSF's award section, the metadata is present in data itself. So, we can directly pull the elements from it.
- c. There could be proper formats like tabular representation, csv etc which will be helpful for extracting the metadata conventions.