

Phenomics Ontology Driven Data Management (PODD)

Overview

Under NCRIS 5.2: Integrated Biological Sciences the commonwealth government has funded two major Phenomics initiatives: The Australian Plant Phenomics Facility (APPF), specialising in phenotyping of crop and model plant species; and the Australian Phenomics Network (APN), which specialises in the phenotyping of mouse models. Both facilities have common requirements to gather and annotate data from both high and low throughput phenotyping devices. The scale of measurement can be from the micro or cellular level, through the level of a single organism, up to (in the case of the APPF) the Macro or field level.

An organism's phenotype, observable and quantifiable traits, is often the product of the organism's genetic makeup, its development stage, disease conditions and its environment. Hence any measurement made against an organism needs to be recorded in the context of these other data. The opportunity exists to create a repository to record the data, its contextual data (metadata) and data classifiers in the form of ontological or structured vocabulary terms. The structured nature of this repository would support manual and autonomous data discovery as well as provide the infrastructure for data based collaborations with domestic and international research institutions. Currently there are no such integrated data management services available to the APPF or the APN.

Note that the original NeAT response was submitted by the APPF only, however the similarity of needs and the relationship of the two facilities means the APN's involvement in this project will both benefit the project by adding to the resources available, and will benefit the APN in resolving their own phenomics data management needs.

Users

This project seeks to support two distinct yet inter-related research communities.

The first is the plant research community, who seek to record, access and manage phenotypic information associated with plant breeding and mutation lines.

The second is the mouse model research community, who similarly seek to record, access and manage phenotypic data about mouse breeding and mutation lines.

For both communities there will be researchers, both national and international, who will want to access and explore the phenomics data and integrate it with other datasets, such as genomic data.

The primary users, i.e. data generators, for both the APPF and the APN would number 50 to 100 per facility per year in the first three years of operation. As the data becomes publicly available it is expected the number of secondary users who would also be accessing to the data would grow quickly and would be in the 1000's by the third year of operation.

The following institutions would participate in this project by setting requirements, contributing resources and steering the project:

- CSIRO Plant Industry: A partner institution in the APPF and host of one of the nodes: the High Resolution Plant Phenomics Centre (HRPPC).
- University of Adelaide: A partner institution in the APPF and host of one of the nodes: The Plant Accelerator (TPA).
- CSIRO Entomology: Lead institution for the development of the Atlas of Living Australia (ALA).

- The University of Qld eResearch Lab: developers of ontology and annotation services and ALA partners
- The Australian National University: A partner institution in the APPF and lead institution for the APN as well as hosting a node of the APN: the Australian Phenomics Facility (APF).
- The University of Melbourne: A partner institute in the APN and the host of one of the APN nodes.

Other institutions and international collaborations that may collaborate in the generation and interpretation of data include:

- CASIMIR (Coordination and Sustainability of International Mouse Informatics Resources), who may provide advice on mouse ontologies.
- TAIR (The Arabidopsis Information Resource) who are interested in data linkage activities.
- Gramene (Data resource for grass research) who may provide advice on plant ontologies.

Needs

There are two key challenges for the phenomics facilities. The first is the ability to provide a Data Management Service (DMS) that can manage data in multiple formats (text, image, video) and not be constrained to a finite set of phenotyping platforms and subsequent data formats. The second is to manage this data and the metadata to support the integration and interlinking of the data within the data management system and provides discovery interfaces to external repositories and services.

Currently no such service is available to these facilities and they are reliant upon saving data to local file servers with no formal specifications guiding the metadata acquisition.

Therefore the needs of the facilities are:

- The provision of a suitable data store which will manage their raw data and analysis results.
- The ability to allow users to easily upload data and metadata to the database.
- The ability to automatically capture data and metadata from instrumentation, where possible. This may require the development of client side software that interfaces with the PODD service level and would be addressed on an as needs basis.
- Support for a range of metadata appropriate to the data generated by the research projects.
- The capture and categorisation of data and metadata about research projects, experiments, materials, biological samples, protocols, data and analysis results to give context for the data beyond that provided by raw data's metadata.
- Interfaces to the metadata to allow for intelligent interrogation of the data resources either by autonomous (i.e. semantic web) or manual methods.
- Security, at the project level, for data that is not declared as publicly available.
- Access and authentication methods that will support trusted electronic communications between the DMS and participating facilities.
- Archival capability for large data files not of primary importance (e.g. original image files).
- The ability to publish data, or make it publicly available, either after a project's conclusion or after a pre-determined period.
- The capability to maintain the data in perpetuity as a valuable research resource.
- The ability to export all data and metadata to facilitate migration to successor applications.

The PODD project will provide a generalised data management solution that targets these needs. Details on how these needs would be addressed are provided in the services section.

The DMS is not expected to provide any analysis software or algorithms for use on any data component. It will not provide any integration with analysis software beyond service interfaces.

Services

The APPF and the APN seek to establish a data management service that would support and add value to their business goals of providing high quality phenomics data generation services. This data management service would include:

- The ability to store data and associated metadata.
- The ability to describe this data according to structured vocabularies and ontologies (this would assist in providing semantic data integration, data mining and reasoning services). This may require the use or development of a vocabulary/ontology registry service that would store ontologies, mappings between ontology terms and mapping between ontology terms and PODD data objects. This would support the interrogation of the data resources through the ontologies registered.
- The ability to describe the data context.
- The ability to archive and retrieve large data files of secondary importance to research. e.g. Image files that have been analysed and are no longer of primary importance.
- The ability to secure research data to protect research intellectual property, share data at the researchers discretion and publish data to the public domain.
- The ability for users to access the data either through a standardised web portal, or through web service interfaces, which in turn would support the autonomous interrogation of the data.
- The ability for users to annotate existing data and projects. This is of particular use when data is publicly available. Researchers who access the data may wish to provide their own annotations against the data. Data owners in turn may wish to respond to this annotation. The annotation in its simplest form would be a series of comments and responses.

This data management service would be maintained and operated by the APPF and the APN.

The overall vision is for users and facility staff to upload project based data in a form that freely allows the interlinking of this data. At the root level is a project entry through which access rights to the data and metadata will be determined. The project is then a workspace for users to upload and interlink/organise their data. For example a user may have information on plant samples which they may upload to the database as records of type: sample. They may then define the experiments that they intend to perform and link individual samples to these experiments. They would also select or define measurement platforms which specify the data generation equipment and the measurement parameters and values. Next the user would load the specific data generated, e.g. a series of images and analysis files. Finally the user may load analysis outcomes, such as text documents that may identify conclusions about the data, or the publication they are preparing.

At any point ontology terms may be used to annotate the information being placed in the project. The project itself may be annotated, the samples or the data. Ontological tags would be embedded in the metadata for each object created in the project.

Information on data entities such as project, samples, experiments, measurement platforms, measurements would be in XML format and thus considered as metadata. The actual data files generated by the phenomics platforms would be the true data. However, it may be that some information, such as sample information, is provided as a data file, e.g. an excel spreadsheet, and only basic metadata captured about the sample file.

The onus is on the user to capture the correct level of information and provide the interlinking. Please refer to the appendix example object relationship diagrams to better understand the process that is being suggested here.

Querying and data discovery is driven by the metadata content. It is not expected that queries will be performed against the raw data. For example a query may wish to discover data in which wheat (defined by the samples) was investigated for wheat rust virus (defined in the experiment) by plant imaging and colour analysis (defined by the measurement or data metadata).

The object classes identified here are examples/suggestions. It is expected that object classes will need more analysis and refinement.

NeAT Characteristics

eResearch effect

Through the use of ontologies to catalogue the data, and metadata to give context to the data, this project will give researchers the capability to store, categorise, annotate, publish and mine phenomic data resources. Researchers will be able to quickly access and analyse data that is of significance as well as investigate relationships in the data through the ontology based index layer.

The project also seeks to encourage the broader research community to access the publicly accessible data through ontology based search and integration methods and therefore allow the linkage of phenomic data resources with other international data resources such as genome databases or tissue banks.

Broader adoption

In the long term this project seeks to establish a generic approach to large scale scientific data management that can be adopted and adapted by other research disciplines with similar needs.

The metadata schemas and the ontology registry service would potentially be applicable across the whole of the eResearch community to facilitate re-use of ontologies and schemas and maximize semantic interoperability.

It is expected some of the specific software tools required by this project would be developed in collaboration with other projects (e.g. the ALA). Some of the specific software tools developed in this project could be made available for reuse by other projects by sourceforge style publication of source code, with an appropriate open source license.

Value adding

Both the APF and APPN will utilise bioinformaticians to not only integrate existing ontologies in the management and annotation of data, but to also recommend to co-ordinating bodies desired extensions to these ontologies.

A review of appropriate backend data management implementations is currently underway with the expectation that the project will be able to incorporate an existing platform to provide the data management component. The current options under consideration are:

- A file based system for management of raw data. An example of this would be the File Access and Management System (FAMS).
- An XML database for the management of XML metadata. This may be in the form of a proprietary database application (e.g. Oracle, MySQL), or a purpose built document repository system such as Fedora Commons.
- An appropriate SPARQL server for management and interrogation of the Ontology registry.

The PODD project would seek the advice of ANDS, ARCS and eResearch groups in the assessment of these applications.

The metadata schema services and repository schema would be expected to be sourced from existing technologies, or its development could be shared with similar projects through ANDS. Some tools for metadata management developed as part of the ARCHER project may be candidates for adaptation and reuse.

The eResearch Lab, at the University of Queensland, has several research activities in the area of ontologies, metadata capture and semantic annotations. This lab would be called upon to provide advice or support development efforts. The participation of this group would potentially significantly reduce development risks.

Web Services based interfacing services and the web presentation layers would mostly be developed internally, with the assistance of groups such as ANDS. The web service based interfaces that are developed may be of benefit to similar projects.

Authentication and authorisation services would expect to be sourced either through ARCS/AAF or through third party software solutions and integrated into this platform.

Standardisation

The use of ontologies is a basic strategy for standardisation amongst biological data domains. Several biological research ontologies are at a mature stage of development and will allow us to adequately describe data resources in terms of: the organism; its growth and developmental stage, its phenotypic traits; and experimental protocols and methodologies.

The APPF are currently seeking out and reviewing international standards for capturing plant research metadata. The APN are likely to engage in a similar activity in the near future. However, the APN has a firmly established connection with CASIMIR who coordinate mouse informatics resources and support collaboration amongst institutes to establish required standards.

A review of appropriate ontologies for the Mouse Phenomics Network is scheduled for late May. For the Plant domain we have identified several candidate ontologies, but will need to determine during the system specification and design phase which ontologies are appropriate to use and do not add too much complexity to the ontological tagging process. These ontologies include:

- Taxonomy – Required for categorisation of organisms under investigation, categorises down to species level, any further categorisation, e.g. ecotype, is not available.
- Gene Ontology – Descriptor of gene function in relation to biological process, molecular function and cellular component. TAIR has links between their gene annotations and GO. Salk has links between their lines and GO. GO:0009644: Response to highlight intensity
- Trait Ontology – Morphological or physiological state or process (RGD pdf) e.g. TO:0000504: leaf temperature
- Phenotype Ontology: Does not exist for plants. Example from MPO is MP:0005533: Increased Body Temperature
- Experiment Ontology – Descriptor of Experiment Goals. Does not exist for plants, however TAIR employ a basic Experimental Method categorisation scheme/ontology. e.g. METH:0000004: Abiotic Treatment
- Measurement Ontology – Measurements applied to determine status of trait. Does not actually exist in any form, but the idea is there.
- Plant Ontology – Descriptor of Plant Structure and also has Plant Growth and Development Stages, e.g. PO:0001053: Leaf Fully Expanded
- Growth Ontology: Probably best accessed using the Plant Ontology, but there is also a Cereal Plant Growth Ontology that can be used: e.g. GRO:0007009: one to two leaves

- Environment Ontology: Descriptor of experimental environmental conditions. e.g. EO:0007075, high light intensity regimen
- Data Integration Ontology; Would be used to provide an indexing service for the data and provide semantic web accessibility to the metadata and subsequent data. The ontology becomes a global reference model for all data in the DMS. It is not yet clear how well defined these ideas are in the biological space.

Where instrument interfacing to the repository is required this will be developed utilising appropriate established methodologies.

Further standardisation will be guided by recommendations from groups such as ANDS.

Key Participants

Reference Group:

- Robert Furbank, Australian Plant Phenomics Facility/CSIRO Plant Industry, Canberra
- Xavier Sirrault, Australian Plant Phenomics Facility/CSIRO Plant Industry, Canberra
- Mark Tester, Australian Plant Phenomics Facility/University of Adelaide.
- Adrienne McKenzie, Australian Phenomics Network/Australian National University.
- TBD, Australian Phenomics Network/University of Melbourne.
- James Eddes, Australian Plant Phenomics Facility/University of Adelaide.
- Jane Hunter, University of Queensland, Brisbane
- Donald Hobern, Atlas of Living Australia, Canberra
- TBD, Representatives from ANDS, ARCS and AAF

Project Committee:

The project would be managed by a steering committee comprised of the following members:

- Adrienne McKenzie, Australian Phenomics Network/Australian National University (Chair).
- The PODD Project Manager (TBD).
- Robert Furbank, Australian Plant Phenomics Facility/CSIRO Plant Industry, Canberra.
- Mark Tester, Australian Plant Phenomics Facility/University of Adelaide.
- Donald Hobern, Atlas of Living Australia/CSIRO Entomology, Canberra.
- Executive Director of ANDS, Ross Wilkinson, or delegate, currently Andrew Treloar
- Executive Director of ARCS, Tony Williams, or delegate, currently Paul Coddington

Project Scale

Project Scale: The project is expected to require up to 5 resources over 2 years. Available NeAT funding is anticipated to be \$250K per annum for 2 years.

Resource Requirements:

- Project Manager, 0.5 EFT
- Lead developer/system architect, 0.5 EFT.
- 2 Software engineers, 1.0 EFT each.
- ALA Software Engineer, 0.25 EFT. (In kind, provided by the ALA)

- Mouse Bioinformatician, 0.75 EFT. (In kind, provided by the APN)
- Plant Bioinformatician, 0.75 EFT. (In kind, provided by the APPF).
- Waiving of overheads by participating eResearch groups (0.5 EFT per software developer and 0.25 each per Project Manager and Lead developer).
- Support and input from ANDS, ARCS and eResearch groups (up to 0.3 EFT)

Neat funding is sought for the roles of Project Manager, Lead Developer/Architect and the two software engineers. Note that the appointment of a project manager and a lead developer is optimistic given the amount of funding earmarked. The PODD project would ask that NeAT consider the need for these two roles in determining the final level of funding.

The cost of facility and hardware requirements are expected to be absorbed by the supporting institutions.

Major Steps

2009 H2

1. Implement data store and a metadata repository onto a development server, set up development environment.
2. Bioinformaticians to develop metadata schemas for the initial set of object classes.
3. Implement a metadata schema repository and an ontology registry for the PODD.
4. Set up basic search service for data discovery and retrieval through ontological terms only.
5. Implement basic load and retrieve services to support loading and retrieval of data blobs and associated meta data.
6. Demonstrate initial system capability to reference group.
7. Secure necessary infrastructure for production version: servers, storage, network fabric.
8. Implement system on production environment for receiving data.
9. Key deliverables for 6 month period: An initial “production” version of the repository with key features in place. “Development” and “Test” versions of the repository.

2010 H1

1. First round user training/familiarisation with system and basic interface.
2. Continue developing metadata schemas to cover all classes of objects.
3. Develop web services to satisfy priority DMS interfacing requirements.
4. Implement access and authentication methodologies.
5. Key deliverables for 6 month period: Web services layer that provides high priority repository interfaces. Acceptable data access and security methodologies. Proven data capture and data access capability.

2010 H2

1. Finalise and develop full set of data handling web services.
2. Extend presentation layer to interface with above services.
3. Extend services and indexing layer and database schema (if necessary) to support semantic web interrogation of data resources.
4. Implement data annotation services.

5. Key deliverables for 6 month period: Complete set of data handling services. Complete indexing and search services. Data annotation services.

2011 H1

1. Extend services layer to support automated capture of data and metadata directly off instrumentation.
2. Implement archiving service for large format files.
3. Implement interfaces with ALA and/or Australian Research Data Commons (ARDC).
4. Refine presentation layer to support ease of use.
5. Key deliverables for 6 month period: Automated data capture and annotation from instrumentation; data archival; final presentation layer (i.e. www interface), external repository interfaces.

Dependencies

Identify dependencies that exist to activities or developments ex project.

The project will have dependencies on the NCRIS AAF project for the provision of federated authentication and access services and the integration with our service.

The integration/extraction of the PODD metadata by the ALA is dependant upon the development of the ALA metadata repository under the DIAS-B project.

The interfacing with the PODD system by the ARDC is dependent upon the progress of this aspect of the ARDC project.

The simple annotation service projected may utilise parts of the annotation service component of the ALA DIAS-B project currently under development, or the annotation technologies being developed for Aus-e-Lit.

The project will also monitor the activities of, interact with and collaborate with the following related national and international initiatives:

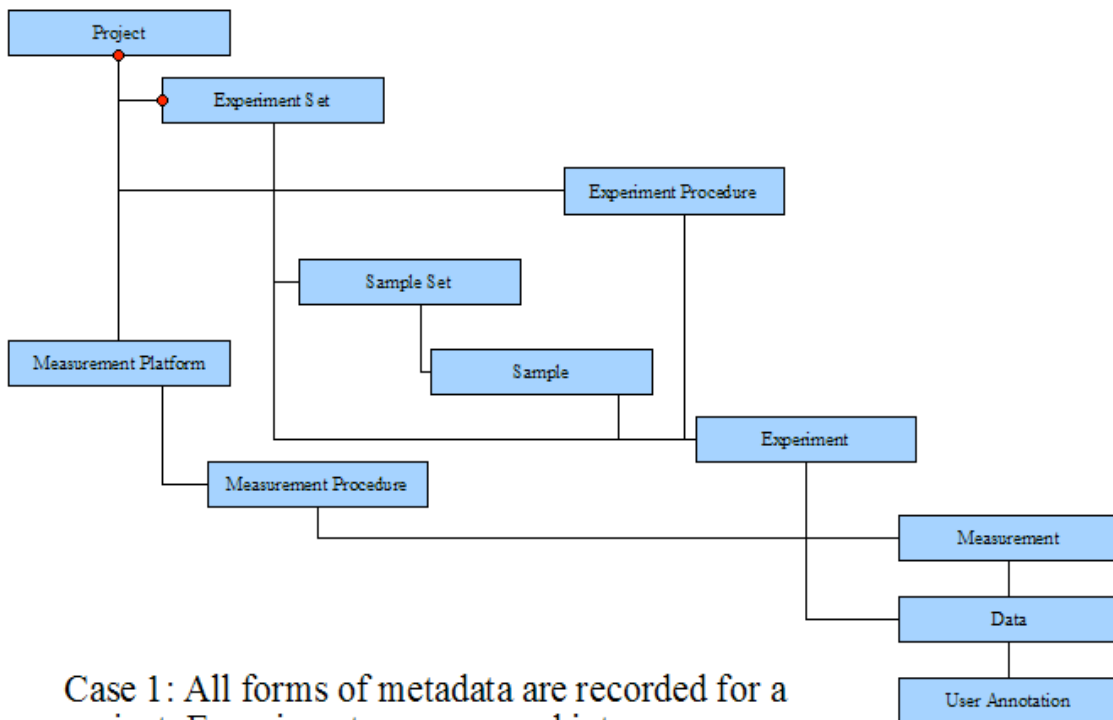
- The International Plant Phenomics Network (IPPN);
- CASIMIR (Coordination and Sustainability of International Mouse Informatics Resources);
and
- The Plant Ontology Consortium.

Appendix: PODD Object Relationship Examples

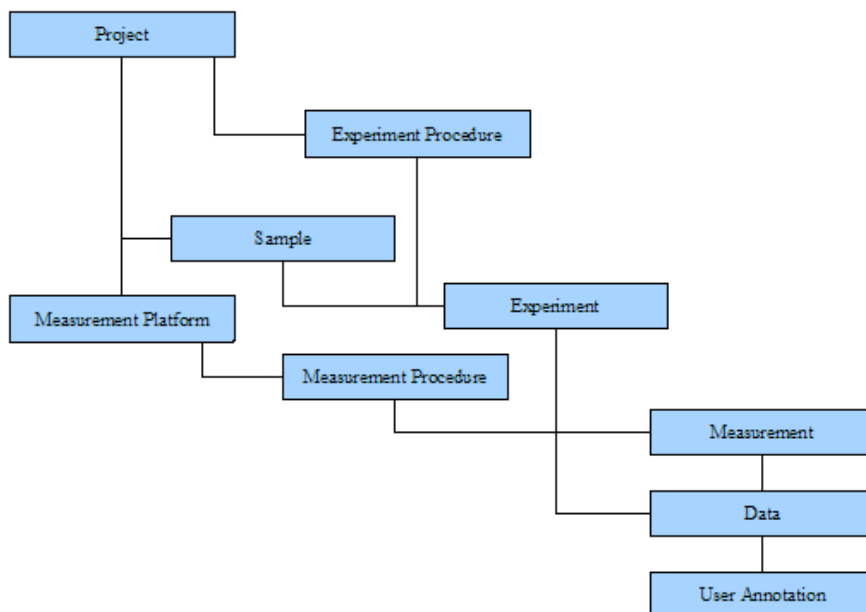
Purpose:

The PODD system should support the notion of managing any type of data, and wrapping it in sufficient metadata to provide context for that data. The possible metadata classes would need to express any type of relationship with any other class. Through these relationships a hierarchy of objects can be maintained, with the Project object at the highest level.

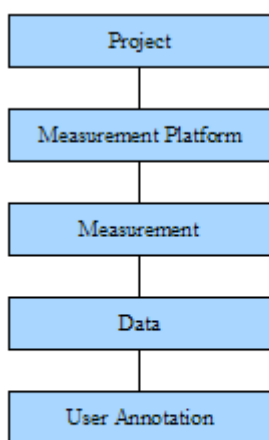
These diagrams attempt to express this idea.



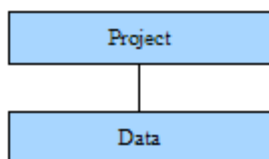
Case 1: All forms of metadata are recorded for a project. Experiments are grouped into Experiment Sets, Samples are grouped into Sample Sets.



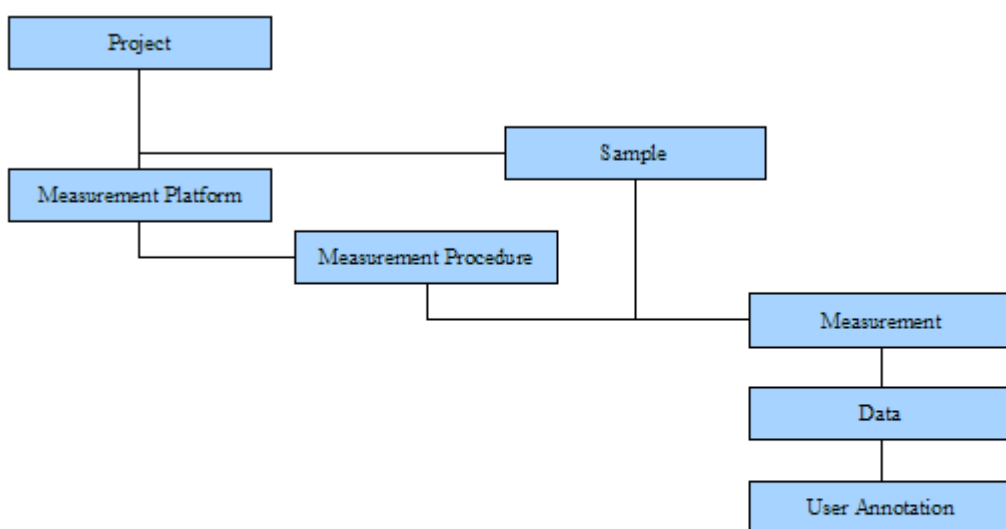
Case 2: Only simple experiments are recorded for a project, no further grouping occurs.



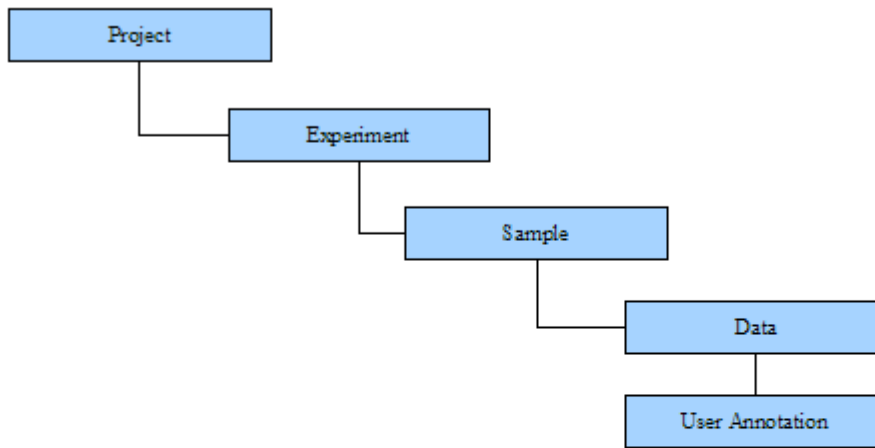
Case 3: Measurements were made and data was recorded.



Case 4: User wishes to place data in repository and provide metadata later.



Case 5: One measurement or lots of measurements about one sample or lots of samples. No experiment information or grouping of samples. Perhaps we are screening samples.



Case 6: No measurement platforms. User may be recording visual observations.