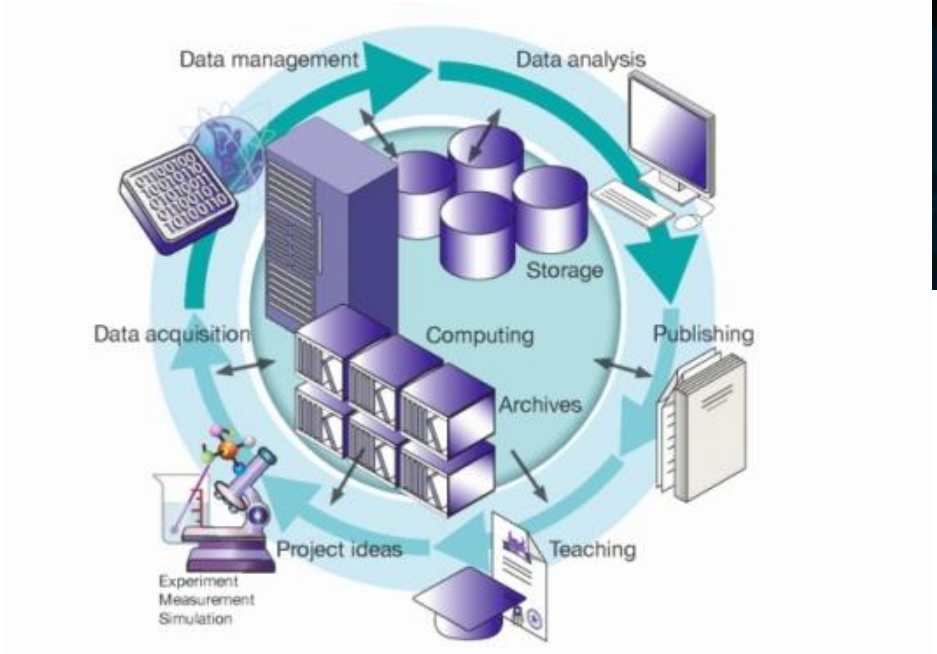
Data Management: the "What", "When" and "How"



Data Management: the "What"…

DAMA(Data Management Association) states that *"Data Resource Management is the development and execution of architectures, policies, practices and procedures that properly manage the full data lifecycle needs of an enterprise."*

A broader and less formal definition (from Wikipedia) states:

"Data management comprises all the disciplines related to managing data as a valuable resource"

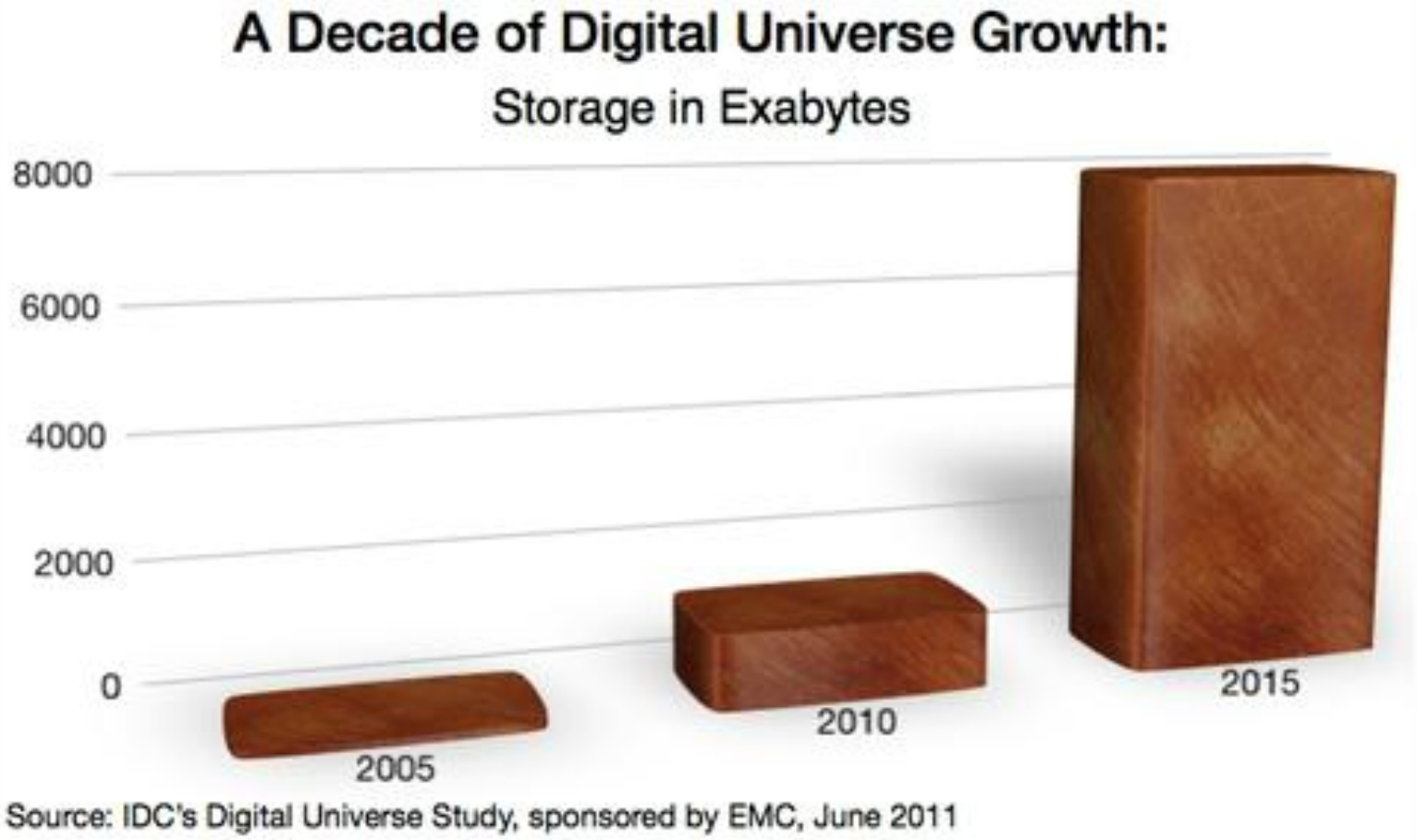
...but What is "Data"?

The word data comes from the plural of the Latin word datum, "**something given**".

Data can take several formats, data is a measurement you gather from your lab, it is a computational output from your computer, it is a lab notebook or even some annotations from your theoretical work... even this presentation is "Data"

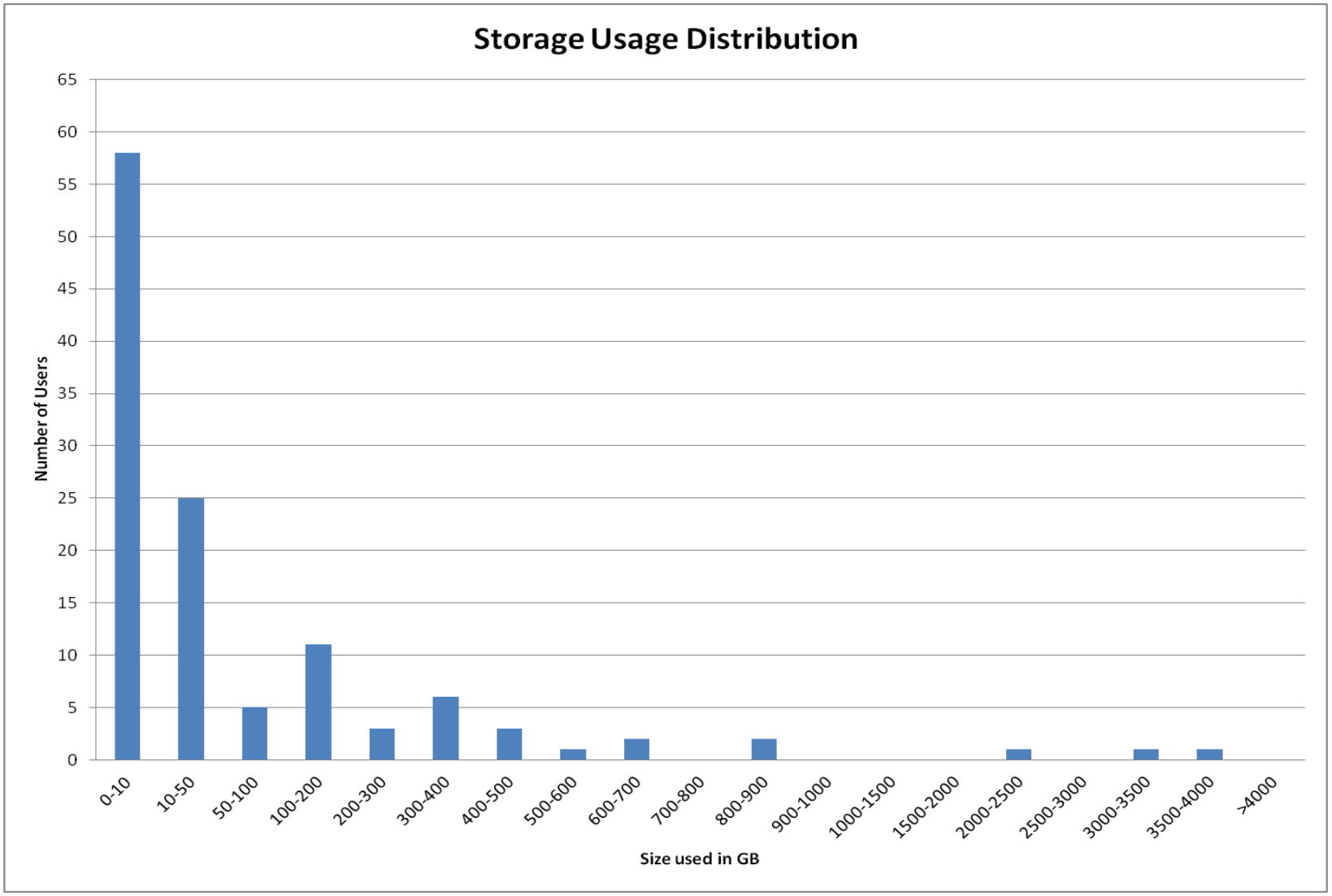
not matter what is the shape of your data you will have to take care of it...or better said, <u>you'll have to manage it!</u>

...what is Data... in numbers?

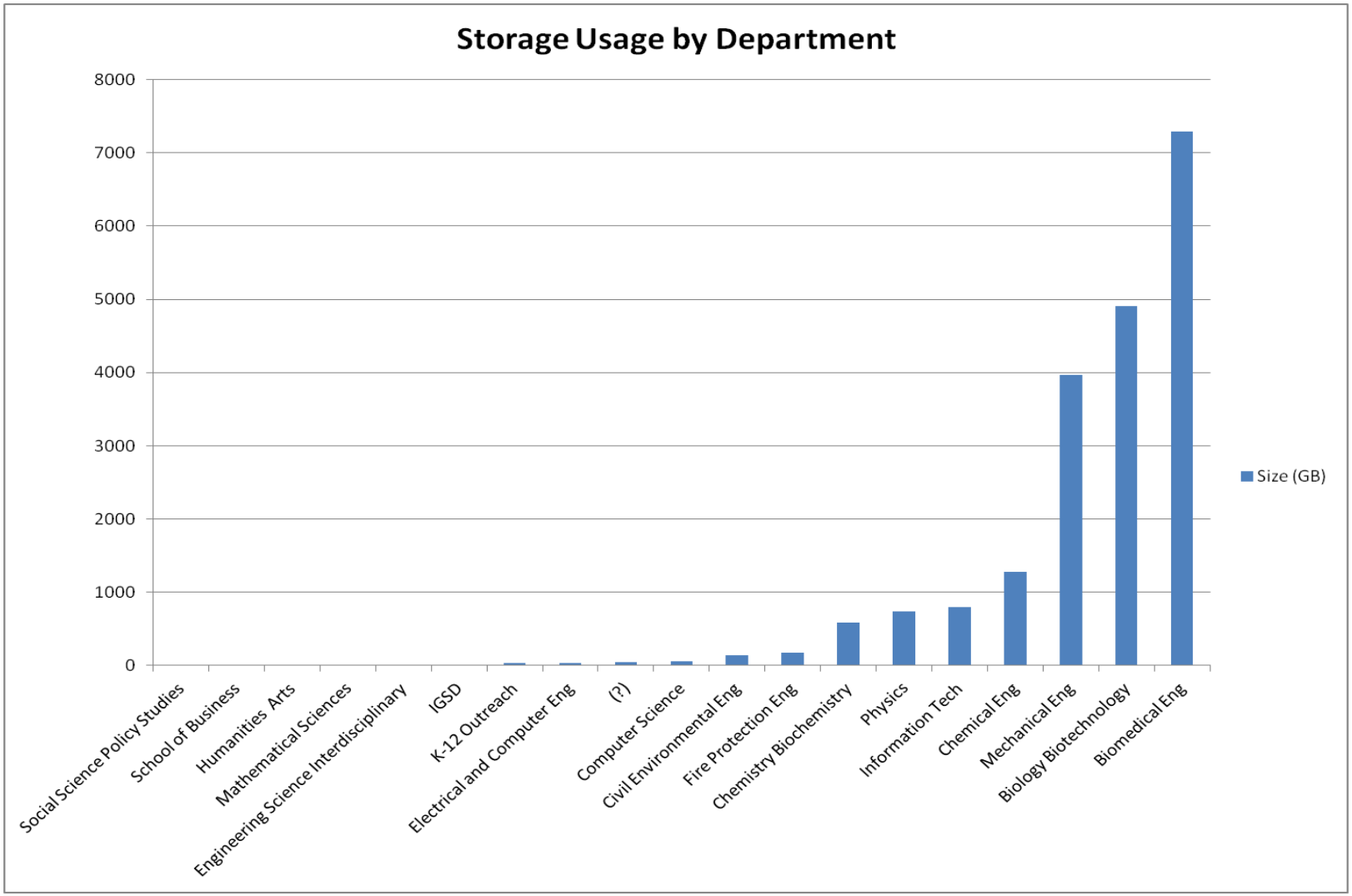


1 Exabyte = 10^3 Petabyte = 10^6 Terabyte = 10^9 Gigabyte

...what is Data... in numbers at WPI?



...what is Data... in numbers at WPI?



WPI Data is growing or exploding?

Data Storage Usage: A brief update on the usage of our research storage servers (\\research.wpi.edu and \\hlemc.wpi.edu).

April 2012
18.2 Terabytes

September 2012
27.7 Terabytes

Usage of storage space has increased by 52% over the last 6 months

Data Management: the "When"…

Three steps to successfully manage your data during their lifetime

STEP1 The Planning: Plan ahead, be aware of

- what type/format of data you will be generating/collecting
- How "large" will be the amount of data collected
- What constraints and restriction will apply to your data (i.e. sensitive data)
- Who will have access to the data produced
- How long you will need to keeping the data
- What are the sharing needs you will have

Data Management: the "When"…

STEP2 The Managing: Take care of your Data while you are generating them, be ready with

- a strategy to organize your data
- a solution to back up your data
- a tool to protect your data from unauthorized access
- a tool that allows versioning of your data (or a strategy that will provide the same functionality)
- a plan to analyze your data on the go and do "housekeeping" when needed (especially important with and expected large amount of dataset)
- a strategy to publish your results

WPI

Data Management: the "When"…

STEP3 The Preservation: Know the future of your data after your research is complete, you should

- determine which data needs to be preserved and for how long
- have a well defined long term access policy
- ensure that your data will be searchable, this is done by assigning metadata to your collections
- Select a repository for long term archiving

Data Management: the "How"…

A good strategy is essential to successfully manage your data, however tools are almost always necessary to implement a good strategy

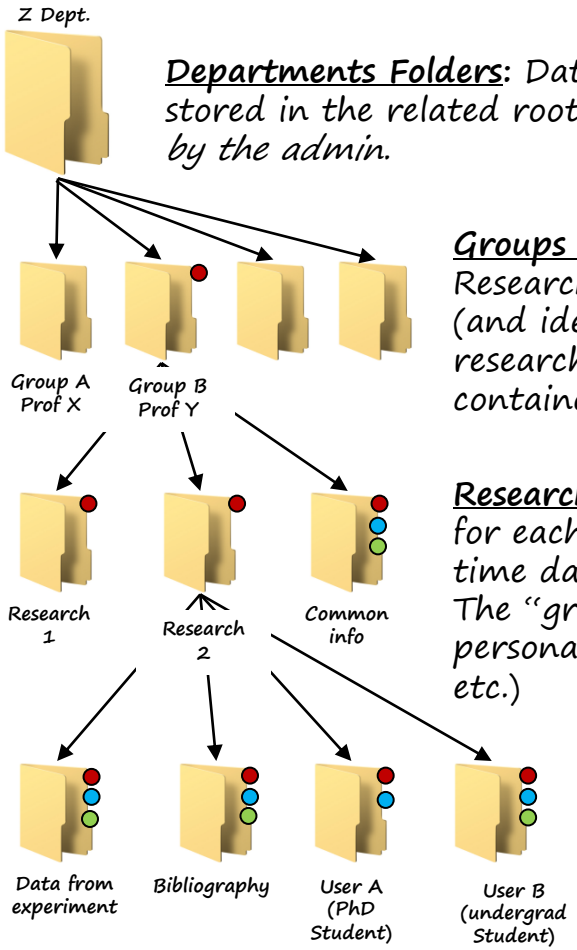
ARCC (Academic Research and Computing Center) is available to help P.I. with creating an optimal data management strategy and identifying potential issues ahead of time.

...how to keep your data organized...

While your research is in progress you will be gathering dataset and it is important to keep them well organized in order to make them searchable and retrievable. Two steps can help you achieving this result:

a) File Naming Convention, in order to quickly identify when a file was created, what it contains, when was last modified, etc.

b) Folder Structure, in order to group efficiently your material, maintain it organized and easy to browse. Folder structures are also key to enforce access policies when required



Departments Folders: Data from each department (AE,BBT,BIO,CE,CH,CS,ME,etc.) is stored in the related root structure. *This is part is invisible to the user and only managed by the admin.*

Groups Folders: Each *research group* should have a proper root folder. Research groups are medium/long term entities, typically managed (and identified) by Professors (or Faculty Personal). The manager of the research group has full control (and visibility) of any subfolder contained within its research group folder.

Research Folders: The "group-folder" level should contain sub structures for each specific research topic. A research group would generate over time data from multiple, independent (or not) research topics. The "group-folder" level should also contain information common to all personal working within the *research-group* (manuals, SOP, calendars, etc.)

Research Related Folders: Data collected from experiment, annotations, bibliography, etc. would be contained within this layer. The structure would reflect specific needs and it would be customized by each involved user. The research personal will have access only to folder related to their work.

Prof Y Ph.D. User A Undergrad User B

...how to protect your data ...

Hardware failure can lead to complete data loss, it is always important to keep a copy of your data. Personal PC, USB memories and other similar devices do not provide a robust backup method for your data. Often data backup for such devices is not automated but left to the user good practice, making the process less reliable.

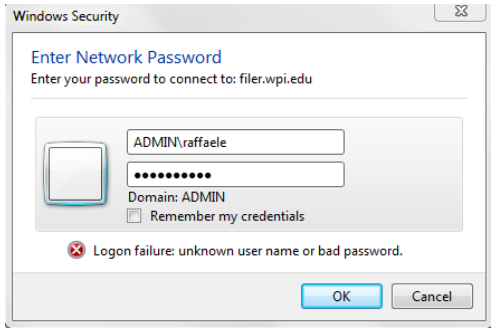
WPI offers several options to safely store your research related data.

- \\research.wpi.edu\ (for faculty and P.I.)
- \\hlemc.wpi.edu\research3\ (for faculty and P.I.)
- \\filer.wpi.edu\ (for everyone with a WPI account)

...how to keep your data safe and remotely accessible...

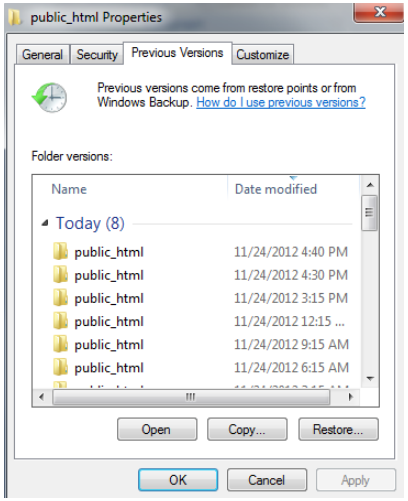
Access to the Data must be granted only to authorized users. Access control is a must for research involving sensitive and personal data but it is strongly recommended for any important stored data. This might be difficult to enforce when using personal devices (laptop, desktop, usb memory, etc.)

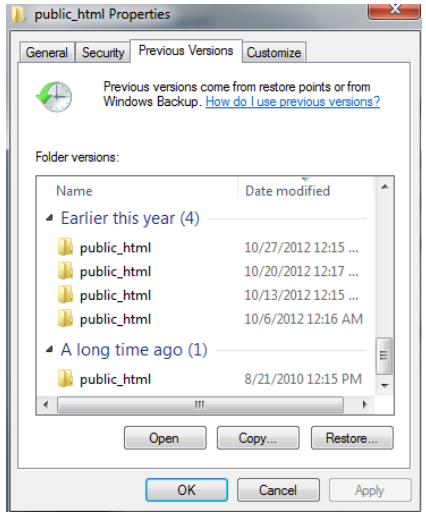
WPI enterprise storage services are password protected and remotely accessible from everywhere.

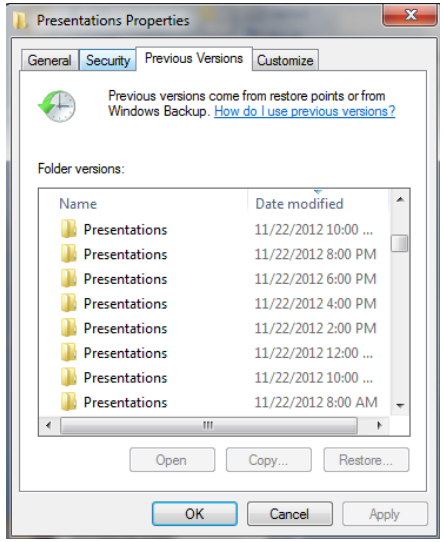


...how to maintain data versioning...

Keep historical copies of documents (and data in general) is a very useful protection against voluntary file alteration or deletion. Versioning can be done manually by keeping multiple copy of a file with an opportune filename convention. WPI storage servers offer to recover data versions for up two weeks with a captured frequency of 2 hours.







...how to share data ...

While your research is in progress you might need to share acquired data with other WPI users or also with users from external institutions. Data on WPI servers can be easily shared within the institution and ARCC is now running pilot programs testing two new powerful tools to share data outside wPI: Oxygen Cloud and iRODS

Many "free" (???) options such as Dropbox, Box, etc are currently available to everyone, however particular care must be used when uploading data on these third party resources.

...how to share data : Oxygen Cloud

What is Oxygen Cloud?

- A tool to share data with other Oxygen Cloud users
- A tool to synchronize data across several machines
- A tool to access data from virtually any platform and any place
- If you are familiar with Dropbox, think of Oxygen Cloud as its twin brother.

What is the difference between Oxygen Cloud and other cloud service provider (Box, Dropbox, etc.)?

- Oxygen Cloud allows us to keep your data stored on a local server machine at WPI
- No copy is made of data during transit through Oxygen Cloud servers
- Data is always encrypted, both on local machines, storage server and during transit between those locations

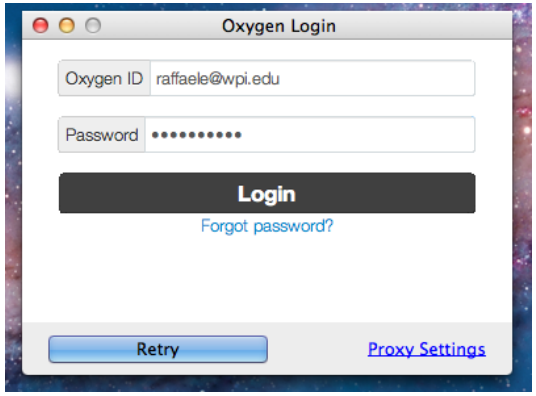
https://oxygencloud.com/

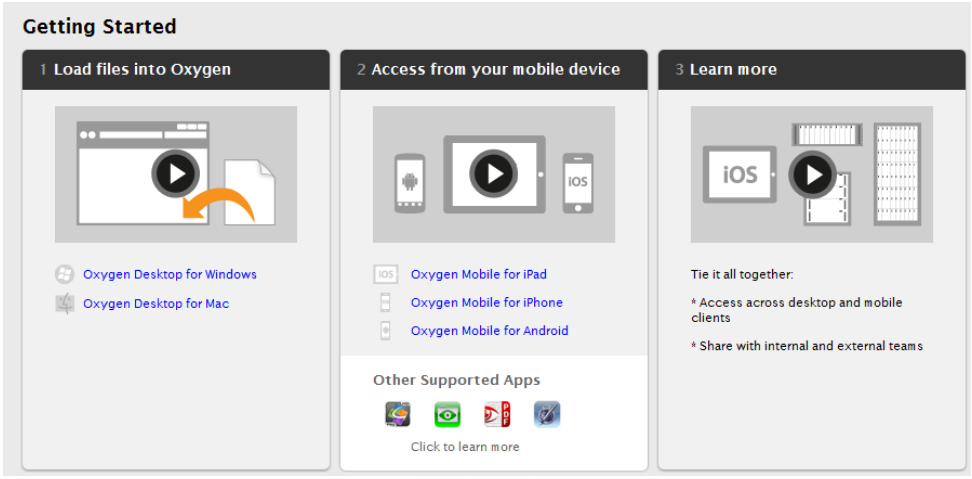
…Oxygen Cloud: how does it work?

- ***Step 1***: Create an account following the link sent in WPI's invitation. The user will get a share on the local storage space at WPI. (Users also receive a 5GB share on the *Oxygen Cloud storage space*, this however is a regular cloud storage space and it is <u>not</u> managed by WPI)
- ***Step 2***: There is no step 2 … just start using Oxygen Cloud

How to access your Oxygen Cloud space:

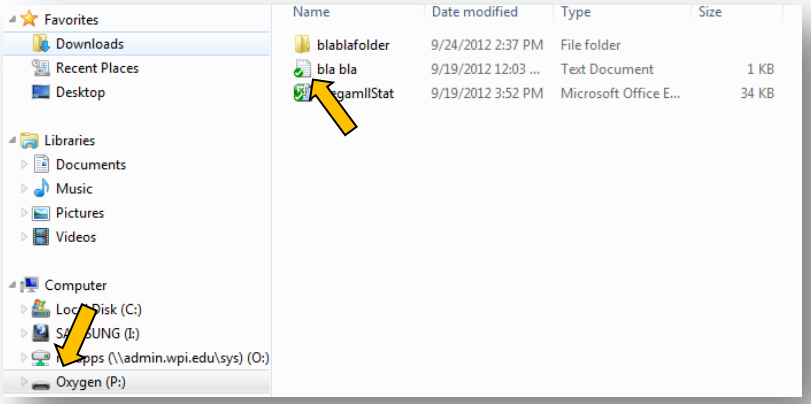
From Windows or Mac machine simply download , install and run the desktop application
From Apple or Android mobile devices, simply install the relative App and lunch it

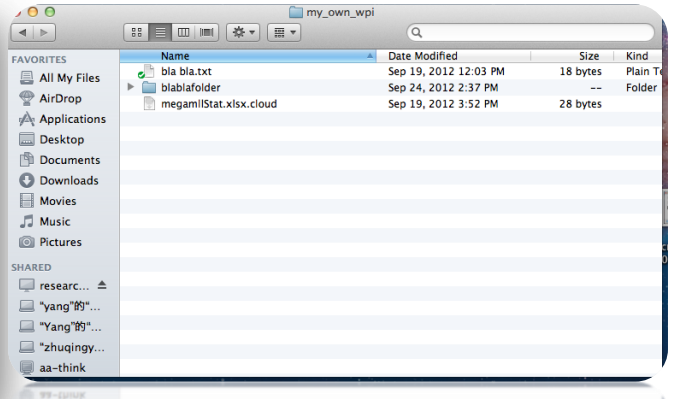


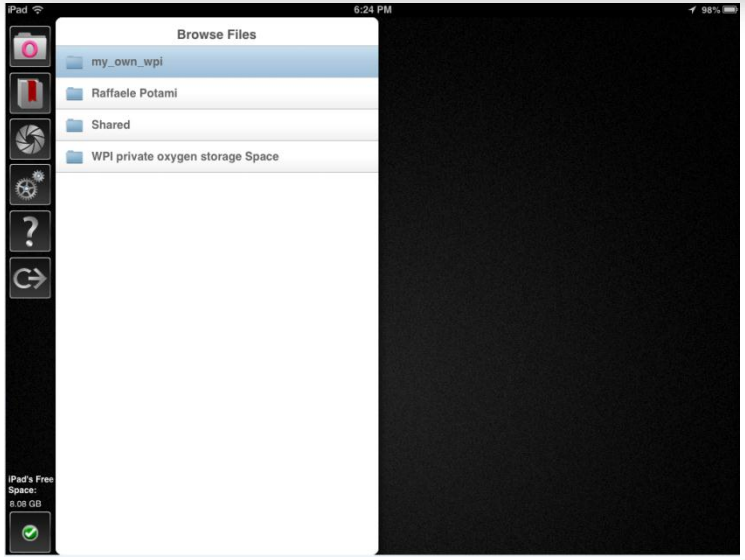


...Oxygen Cloud: how does it look like?

On Windows and Mac machines Oxygen Cloud is mounted as an additional drive and it can be used similarly to any other folder.







iPad and iPhone interfaces are quite smooth and easy to navigate making the data accessible also from mobile devices.

...how to share and organize data : iRODS

- iRODS stands for Integrated Rule Oriented Data System
- It is an open source data grid designed by DICE group from University of North Carolina at Chapel Hill to help people organize and manage large collections of distributed digital data
- It allows users to share organized collection of data
- It is a highly scalable data managing tool which allows the use of metadata to easily categorize, search and retrieve personal and shared data
- It is based on the concept of "zones", WPI would constitutes a zone while multiple zones can be linked into a "federation" . This strategy is intended to facilitate data sharing between different institutions (zones), each running its own version of iRODS.

www.irods.org

...iRODS: how does it work?

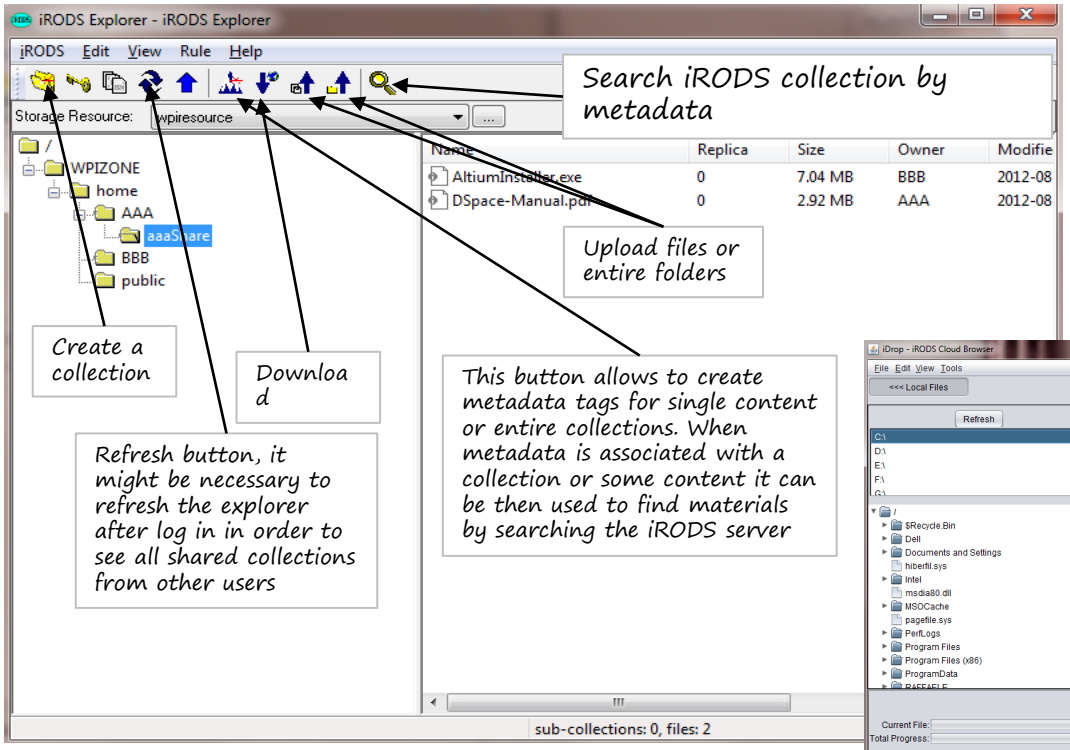
iRODS software allows users to upload, download and manage data on a storage server located at WPI. Each user has a main folder/collection and can create sub-folders or sub-collections as well as upload individual data. Users can share collections and their content.
Users accounts are (at the moment) independent from WPI credentials and will be provided to WPI faculty and research personal as well as external individual when requested by WPI personal.

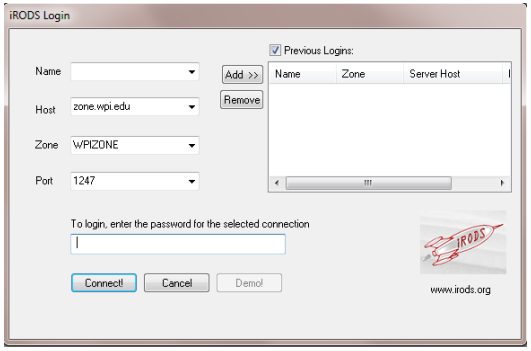
iRODS server at WPI can be accessed in different ways:

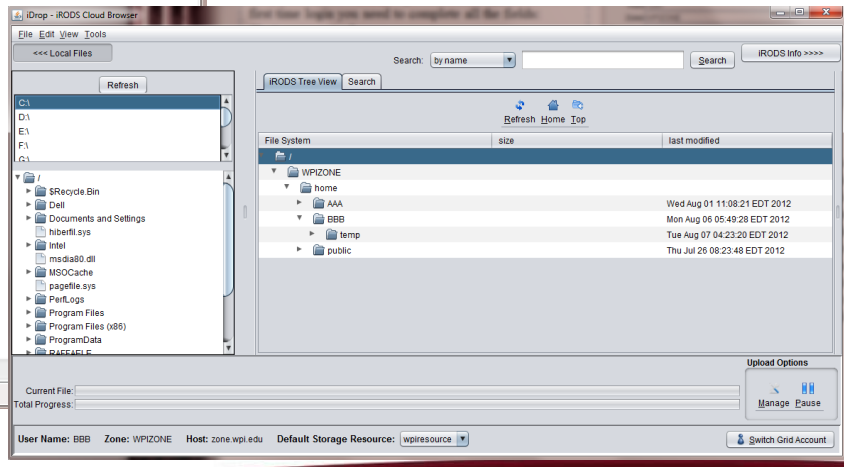
- Windows Explorer Client (only for Windows OS, can be downloaded from http://zone.wpi.edu/clients)
- Web Browser Client (any OS, access from http://zone.wpi.edu/wpizone)
- Java Explorer Client (any OS once Java is installed, can be downloaded from http://zone.wpi.edu/clients)
- iCommands (Windows, Unix/Linux and Mac)

...iRODS: how does it look like?

The iRODS explorer will appear, its functionality are quite similar to the standard windows explorer. The functionalities of the iRODS are quite intuitive, most operations can be performed by selecting an object and clicking the mouse right button. The explorer support drag and drop of files or entire folders.

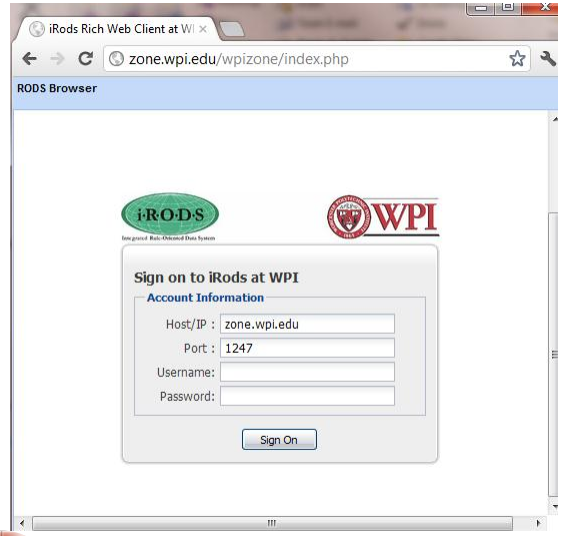


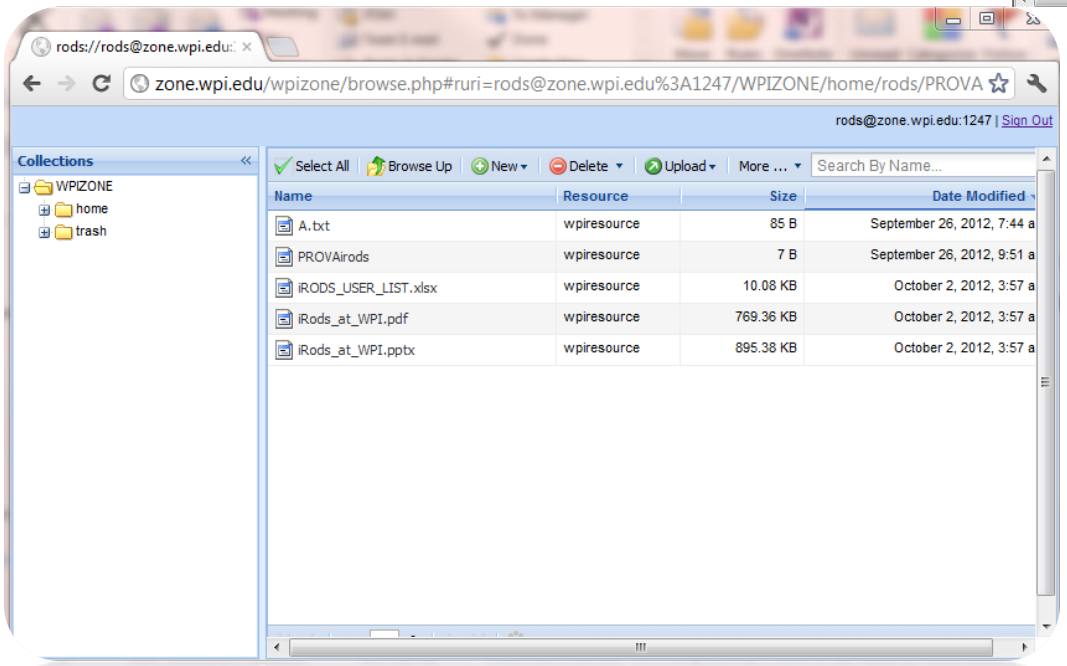




...iRODS: how does it look like?

The web client is an easy and powerful way of accessing data collection. The user can navigate into collections by clicking on its name and up by using the [Browse Up] button. The web client allows to add and search by metadata and support multiple files upload via drag and drop capability.





...how to share and organize data : SharePoint at WPI

Sharepoint is a service offered to members of WPI Community to provide a collaboration environment that is tailored to online document collaboration.

It offers:

- Office Web Apps: Edit a document from within a browser
- Collaborative, real-time document editing
- New Lists/Library types
- Browser support (IE, Chrome, Firefox and Safari)

https://sharepoint.wpi.edu/Pages/Default.aspx

https://www.wpi.edu/Academics/CCC/Help/Software/sharepoint.html

...how to keep "paper-like" data organized...

How to manage non digital based data, such as papers or lab notebook? Similar idea of what discussed still applies...or you can convert your paper based data to a digital format.

There are also tools to transform your paper based data in a digital format as soon as you get them: use a smart pen http://www.livescribe.com/en-us/smartpen/

Now your data will be secure, protected and well organized!

Questions?