SENG8080-23F
BIG DATA CASE STUDIES

Lecture 2

Agenda

- Last Week’s Study
- Project
- Collaboration
- Podcasts
- Research

CASE STUDY: NFL

Six Elements of Any Big Data Project

Data Research and Integration

This is focused on provenance and integration points of data.

Data Collection

Data Collection is normally an ongoing process of extracting and collecting data; often this is at regular intervals or in real time.

Data Storage and Maintenance

Data needs to be stored somewhere, not just today but for the future. Some considerations here are regulatory and some are practical (i.e. how long to store the data).

Data Quality

Data comes from all sources and can be accurate or reliable to varying degrees. Data quality tried to track and maintain this.

Data Analysis and Visualization

This is where the value comes in. Why are we gathering and keeping this data?

DevOps

Integrates with existing systems to speed the development cycle.

What are Case Studies and Why Look at them?

- Descriptions of real-life problems and how practitioners attempted to resolve them.
- Evaluation of what works and what does not work.
- Opportunity to replicate the efforts of others to practice them.

https://www.youtube.com/watch?v=ypbSMS8XrAE

Example Case Study

Consider the following case study:

https://www.datapine.com/blog/the-power-of-big-data-in-american-football/

In groups, covering each of the six elements, evaluate and be prepared to report on:

a) Description of what was done

b) What worked? What didn't?

c) Is the information provided in the description credible?

d) In what other areas could this be applied?

PROJECT

Report and Presentation

- Class divided into groups of 3 or 4 students
- Assignments
  - Each group will take on a project.
  - To do: Research project, write report, present/demo project to class.
  - Deliverables/Check Points
    - Project Team: **End of Week 1**
    - Project proposal: **End of Week 3**
    - Mid-Term Progress: **End of Week 7**
    - Presentation/demo and report: **Last 2 or 3 classes (depending on group sizes)**
  - Group work, individual assessment.

COLLABORATION

Version Control or
“Umm, is this latest? Or was it…?”

- Version Control/Revision Control = How do we manage and keep track of changes to set of files, procedures, prototypes, etc.?
- Version Control/Revision Control System = Combination of tools and processes that executes VC and allows for reversion to previous revisions.
- Generally, can:
  - Retain and produce history of documents and files.
  - An internal tagging system to track updates, etc.
  - Can be cloned efficiently as needed.
  - Self-contained (i.e. needs no other software or applications).

What do all these words mean?

Trunk:

- The main ‘branch’ of development.

Branches

- Just like a tree, when two development paths diverge (at a *fork*) it is helpful to track it.

Repository

- The central “place” where team members store all files related to the project.

Tags

- Give meaningful descriptions to stages of development. E.g. Final_Pres instead of V.251.12.32a(1.5e)

Working copy

- The version that is currently being worked on.

Commit changes

- Moving a version from the ‘working copy’ to the permanent revision.

Three flavours of Version Control

Local

- Resides on individual machines

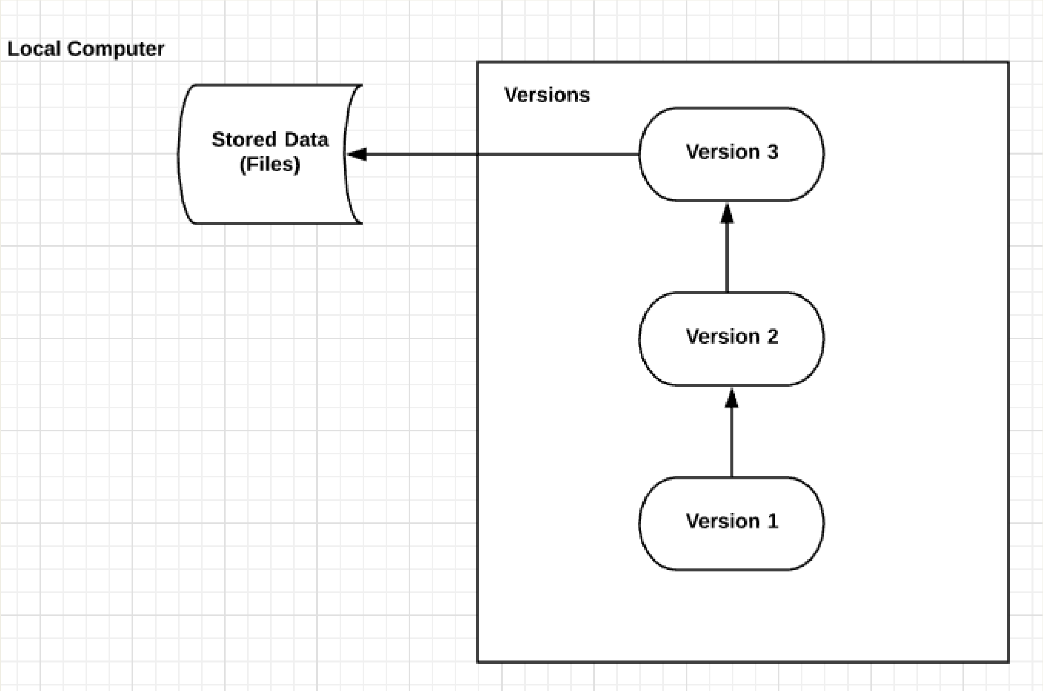
Centralized

- Co-ordinated on one centralized server.

Distributed

- Clients 100% clone the repository

Local Revision Control System



Edonix, 2018

Advantages

- Easy
- Common
- Most tools (E.g. Word, Excel) have a built in RCS.

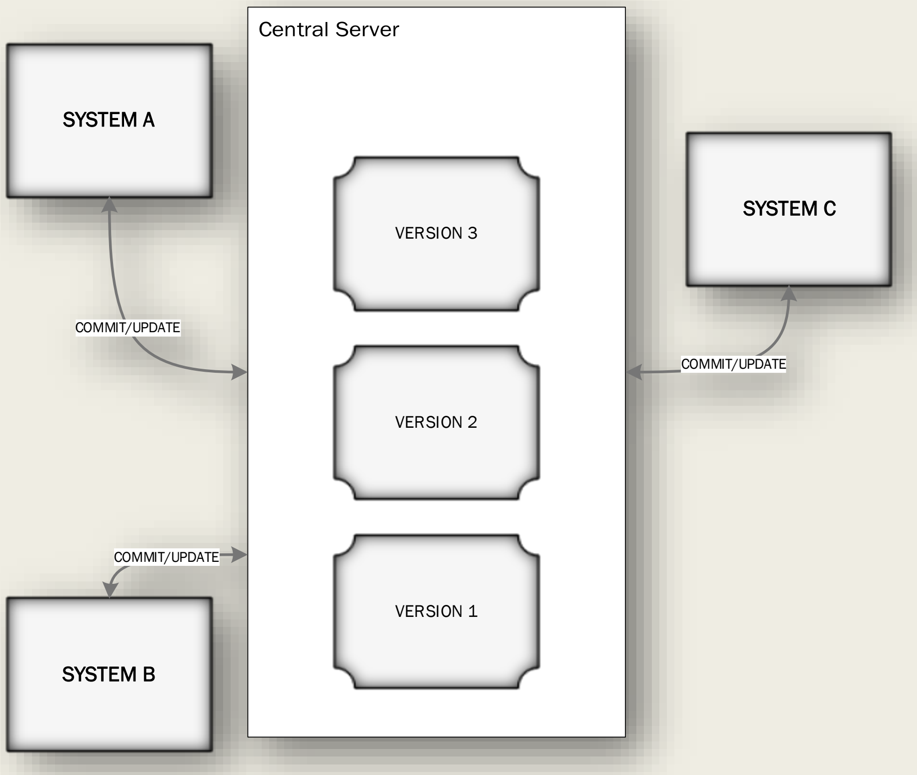
Disadvantages

- Error Prone
- No redundancy for failure

Example

- Word, Excel

Centralized Revision Control System



Advantages

- Redundancy
- Tighter Controls
- Tracks clients
- Most tools (E.g. Word, Excel) have a built in RCS.

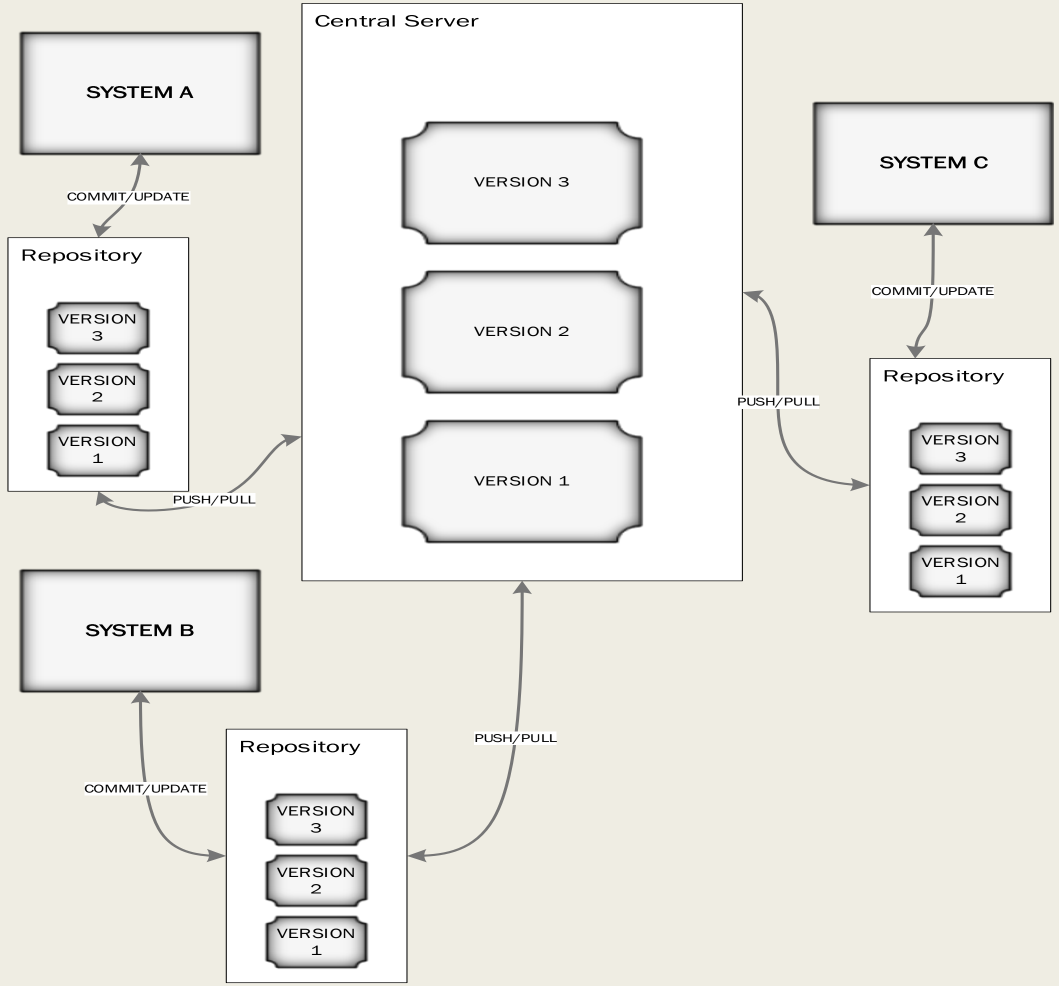
Disadvantages

- Back-Ups
- Communications Dependant

Examples

- Tortoise SVN

Distributed Revision Control System (Simplified)



Advantages

- All of the Advantages of Centralized
- Complete clones exist in multiple locations (i.e. backup!)
- Scalable

Disadvantages

- Cost
- Can be complex to maintain

Example

- Git, Mercurial

What problem(s) does Git solve?

- Coordinating software development (or anything that requires collaborative version control)
- Multiple users working on a codebase/project
  - *Decentralized collaborative development*
- Files need to be tracked and synchronized
  - *History of changes to each file*
  - *Merge conflicts*
- Reverting changes
- Diffs between versions

Let’s Look at Git

- Used by > 80% of open source developers
- A quick video from Git:
  - *What is VCS Git SCM • Git Basics #1 – YouTube*
  - *Do NOT get lost in the details. Concentrate on the Big Ideas for Git and GitHub.*
- What is GitHub? A repository hosting service for Git.
  - *https://www.youtube.com/watch?v=w3jLJU7DT5E&t=102s*

Examples

- Look at Apache Beam and see if you can answer the following questions
  - *https://github.com/apache/beam*
  - *Who are the top contributors?*
  - *How do we contribute?*
  - *What are some top bugs?*

- Assignment 1 will give you an introduction to how to use GitHub

Useful Git References

- Starting tutorial: https://guides.github.com/activities/hello-world/
- “The” book: https://git-scm.com/book/en/v2
- Git cheatsheet: https://www.atlassian.com/git/tutorials/atlassian-git-cheatsheet
- Github learning: https://lab.github.com/

PODCASTS

RESEARCH METHODS

FINDING YOUR SOURCES OF MATERIAL

Refine your topic

- Some sources that can help refine the topic:
  - *Library catalogues, indices of books, databases, encyclopedias, newspapers, or magazine articles*
  - ***Conestoga College online Library***
- Example: Library catalogue (e.g. Library of Congress):
  - *A search for “Big Data" returns:*
    - 1,459 items in the current collection
    - Subjects including: Innovation in Big Data, Modelling Big Data, Understanding Big Data, NoSQL, and many, many more.

Plan the Information Search

- If possible, identify the “big ideas” of your topic.
- Carefully choose keywords for your searchs using Big ideas *and* synonyms

- Develop search terms:
  1. *Begin with one search item: likely the results will be too big.*
  2. *Add a additional search terms using Boolean operators*
     1. AND will narrow a search
     2. OR will widen a search

Determine Sources

Sources to use for an overview of a topic:

- General encyclopedias, e.g. Encyclopedia Americana®
- Specialized encyclopedias, e.g. Encyclopedia of Drama®
- Specialized dictionaries. e.g. Dictionary of Classical Mythology®
- General interest periodicals (magazines and newspapers)

Sources to use for specialized information:

- Books (search the online catalog for specific titles; browse call number location for related titles)
- Scholarly journals (journals published for academics or professionals; the library's holdings include print journals and electronic databases with full text)
- Internet (use caution with this source; you must evaluate Internet sources for credibility, authority, and currency)

Keep track of sources used for citing documentation:

- Write down all the publication information, pages used, etc.
- Know what style format is required for citations (example: APA)

With your research completed, you are ready to start writing your paper!!

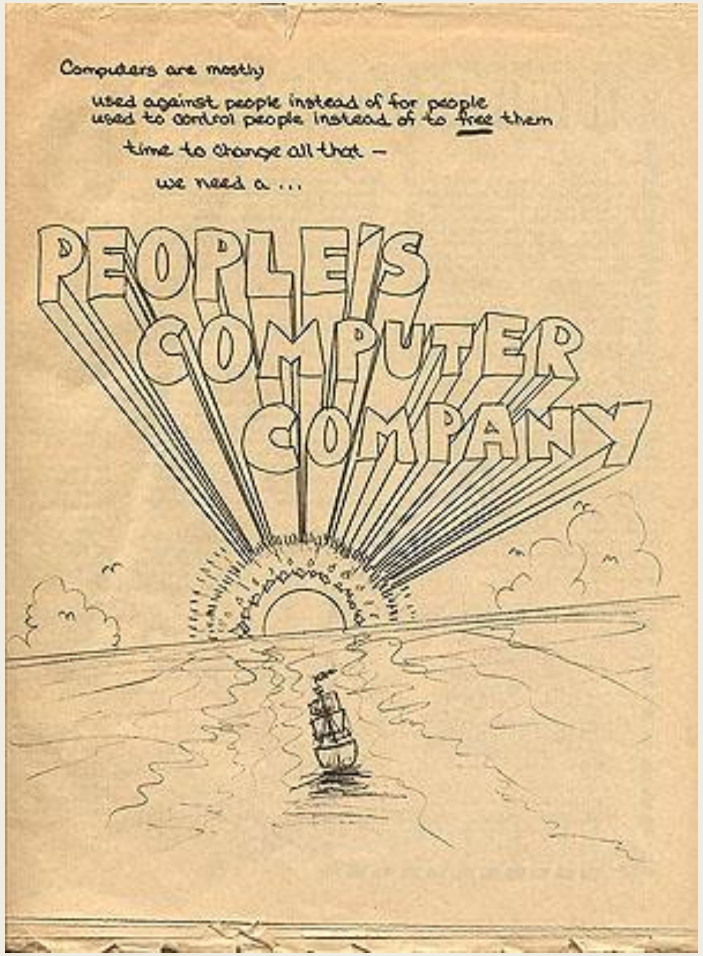
CATEGORIZING SOURCES

With resources from dr. Steve beatty, and the university of Washington

Types of Sources

1. Primary Sources

Primary Sources



Computers are mostly

used against people instead of for people
used to control people instead of to free them

time to change all that –

we need a ...

PEOPLE'S COMPUTER COMPANY

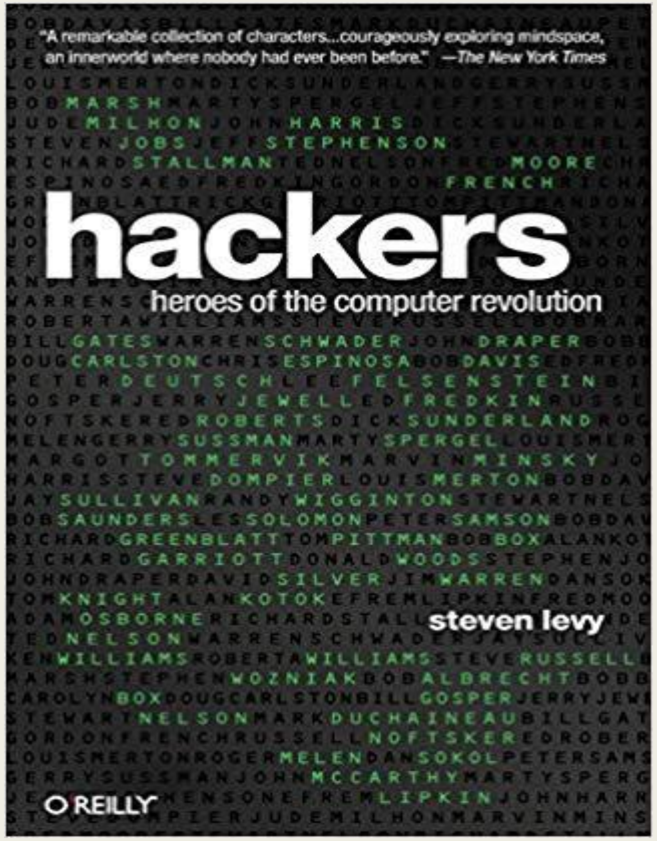
- A first hand document

  For example, when looking at the history of personal computing and specifically the People' Computer Company's contribution to it's growth, issues of the PCC newspaper would be a primary source.

Types of Sources

1. Primary Sources
2. Secondary Sources

Secondary Sources


"A remarkable collection of characters...courageously exploring mindspace, an innerworld where nobody had ever been before." —The New York Times
hackers
heroes of the computer revolution
steven levy
O'REILLY

- An indirect account.

For example, when looking at the history of personal computing and specifically the People' Computer Company's contribution to it's growth, the book *Hackers,* (specifically Chapter 8) be a secondary source.

Types of Sources

1. Primary Sources
2. Secondary Sources
3. Scholarly Sources

Scholarly Sources

- The easiest way to identify a scholarly source is to see if they document their sources. For example, *MacLeans* does not include parenthetical documentation or a Work Cited after its articles.

  Why? Because it is written for a mass audience not a scholarly one.

Scholarly Source

- *MacLeans*? No.
- *The Journal of Information Technology Research*? Yes.
- *The Journal of Big Data*? Yes.
- *Chatelaine*? No.
- *The Globe and Mail*? No.
- *My friend who works at Google.* No.

Newspaper Hierarchy

Top 6 Canadian/American Newspapers

- *The New York Times*
- *The L.A. Times*
- *The Washington Post*
- *The Wall Street Journal*
- *The National Post*
- *The Globe and Mail*

Newspapers

- There is a hierarchy of newspapers. You would generally use the papers on the previous slide before *The Waterloo Chronicle*, for example.
- When wouldn't you? If it were a local topic, like start-up culture in the Waterloo Region, where *The Waterloo Chronicle* might be seen as more authoritative.

Source Criteria

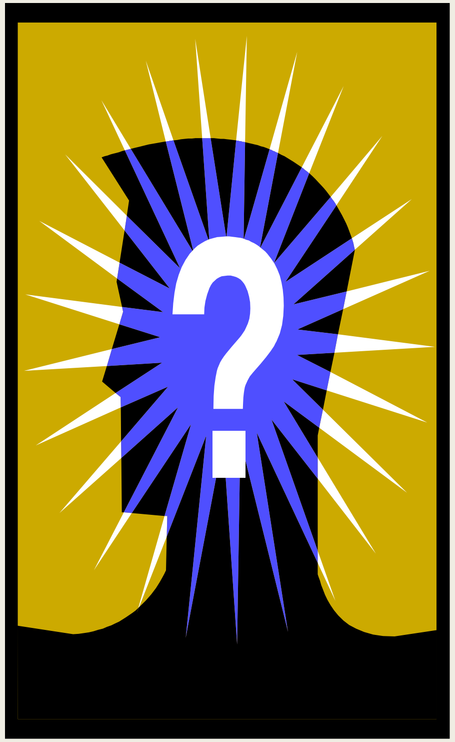
- You’ve run a search and found 25 possible sources. So how do you decide which sources are the best sources?
- First criterion?

Source Criteria

- Authority

If you’re researching exploratory data analysis, this guy is an authority on the subject.

Source Criteria



- What if you don’t know who the authority is?

- Review the literature and see who is constantly cited. That is, check the indices and bibliographies.

Source Criteria

Gartner®

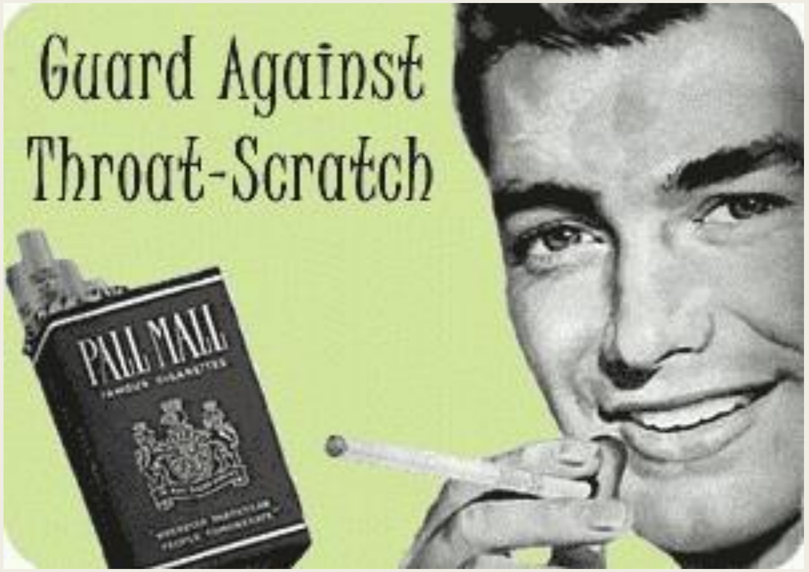
- Sometimes, the authority is a government agency (for example, in post quantum cryptography).
- Or it could be a recognized industry thought leader (e.g. a White Paper from Gartner).

Source Criteria

- Currency

In general, more current sources are preferred. For example, data backup is quite different now (e.g. virtual machine backup) than it was in 1975.

Source Criteria



- Objectivity

  In the 50s, cigarette companies sponsored studies that claimed smoking helped reduce tension and sore throats.

Of course, that would never happen today….

COMPLIANCE TRAINING:

facebook

WHY DATA PRIVACY MATTERS

Written by Jim Bachert

Director, Compliance Learning - West Coast US, Interactive Services

interactive
compliance training

SHOULDN'T I TELL YOU ALL ABOUT ME?
NOT NECESSARY. JUST SHOW ME YOUR FACEBOOK ADS!
ONCE MORE MR. FREUD RECOGNIZED IN THE REALLY EARLY DAYS OF WEB 2.0 ITS VALUE IN PSYCHOANALYSIS
geek and poke

Source Criteria

Authority

- Is the source reliable and comprehensive?

Currency

- How current is the source of information?

Objectivity

- Is the source trustworthy or does it have mixed motives?

Categorize these sources for “Big Data”

1. Level of Authority
2. Currency
3. Objectivity

- *MacLeans?*
- *The Journal of Information Technology Research?*
- *The Journal of Big Data?*
- *The Future of Technology, by Tom Standage, published 2005*
- *Quora?*
- *Chatelaine?*
- *The Globe and Mail?*
- *My friend who works at Google.*

More information…

- ***Research 101:*** *(http://faculty.washington.edu/jwholmes/research101/index.html)*
  - Available from the University of Washington. A terrific introduction to all aspects of research. Choose only the ones that are relevant but remember this source for future use!
- ***Research:*** *http://www.ipl.org*
  - Part of the Internet Public Library this contains a lot of wonderful information on many subjects, including research.