LTI
Language, Technology and the Internet

Collaboration and Wikis

Francis Bond
Division of Linguistics and Multilingual Studies
http://www3.ntu.edu.sg/home/fcbond/
bond@ieee.org

Lecture 5

Revision of Email; Usenet; Chat and Blog

- All share some characteristics of speech and text
- Usage norms not fixed
- Communication methods may disappear before the norms are fixed (Usenet)
- Large scale discourse studies still to be done
- Some genuinely new things
  - time-lagged, multi-person conversation
  - un-edited text

Email

Speech like	Text like
time bound*	space bound (deletable)
spontaneous*	contrived*
face-to-face	visually decontextualized
loosely structured*	elaborately structured*
socially interactive*	factually communicative
immediately revisable	repeatedly revisable*
prosodically rich	graphically rich *

Usenet (asynchronous)

Speech like	Text like
time bound*	space bound
spontaneous*	contrived*
face-to-face	visually decontextualized
loosely structured*	elaborately structured
socially interactive*	factually communicative
immediately revisable	repeatedly revisable
prosodically rich	graphically rich

Chat (synchronous)

Speech like	Text like
time bound*	space bound
spontaneous*	contrived
face-to-face	visually decontextualized
loosely structured*	elaborately structured
socially interactive*	factually communicative
immediately revisable	repeatedly revisable
prosodically rich	graphically rich

Blogs

Speech like	Text like
time bound	space bound
spontaneous*	contrived*
face-to-face	visually decontextualized
loosely structured*	elaborately structured*
socially interactive^{comments}	factually communicative
immediately revisable	repeatedly revisable*
prosodically rich	graphically rich *

Collaboration and Wikis

Overview

- Version Control Systems
- Wikipedia
- Some issues within Wikipedia
- Licensing and Ownership

Versions, Revisions, Authorship

- Many works have multiple versions
- Before writing, every production was different
- With writing, every copy was different (before printing)
- The same source may have multiple translations
- Authors (and Editors and Publishers) revise text
  - Examples?
- Computers can store multiple versions together

Revision Control Systems

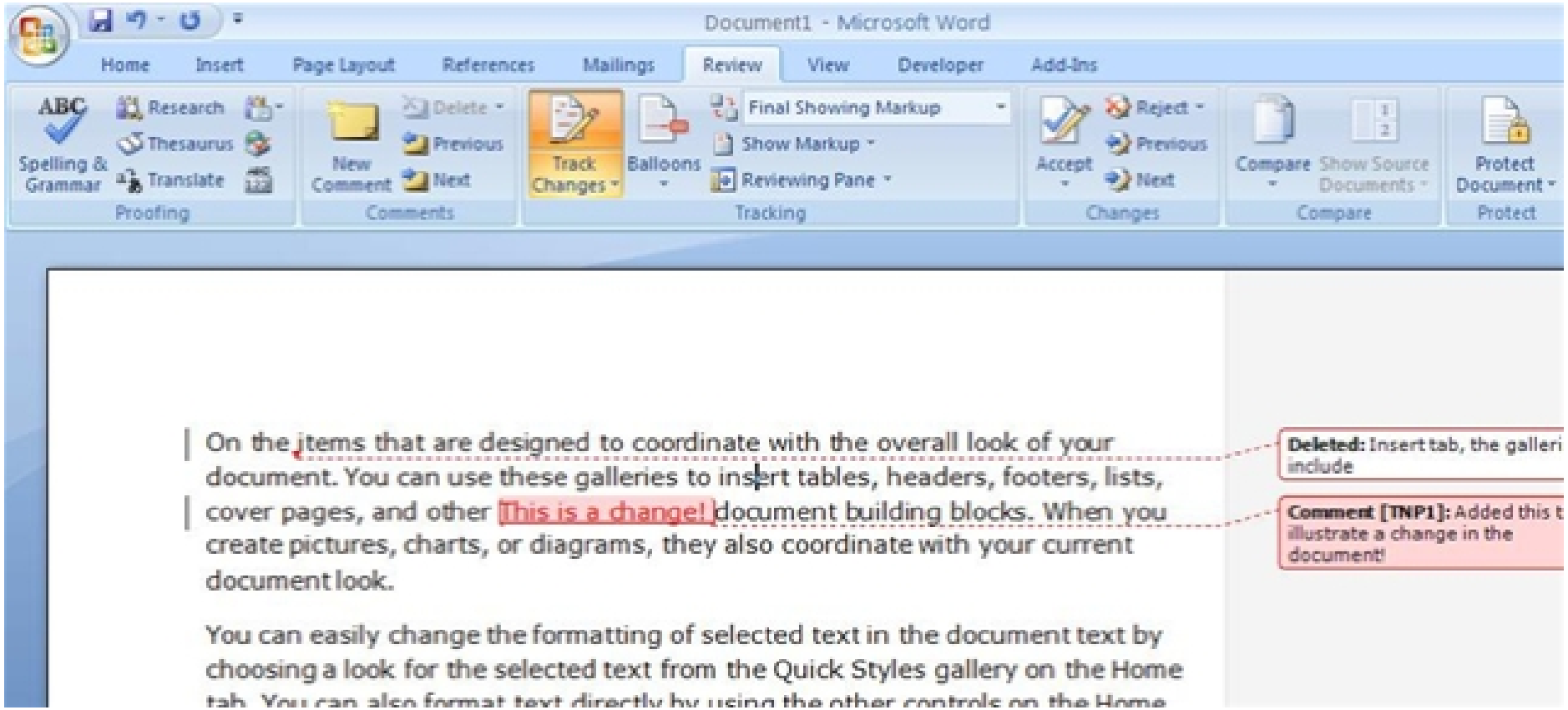
- Versioning file systems
  - every time a file is opened, a new copy is stored
- CVS (Concurrent Versioning System), Subversion, Git
  - changes to a collection of files are tracked
  - simultaneous changes are merged
  - github is a popular interface: github.com/fcbond, github.com/bond-lab
- Revision Tracking
  - Revisions are stored within a file (MS Word, Google Docs)
- Authorship in shared writing (who wrote what?)

svn blame

```
  92     siegel
   2     siegel ;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
   2     siegel ;;;      file: japgram
   2     siegel ;;;  written by: Melanie Siegel/Emily Bender
   2     siegel ;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
   2     siegel ;;; author            | date         | modification
   2     siegel ;;; -----------------|-------------|-----------------------------------------
  94       bond ;;;Melanie Siegel (MS)|            | Emily Bender (ERB), Francis Bond (FCB),
  94       bond ;;;                  |             | Chikara Hashimoto (CH),
 421 michael.goodman ;;;                  |             | Takaaki Tanaka (TT), Akira Ohtani (AO),
 421 michael.goodman ;;;                  |             | Michael Goodman(MWG)
   2     siegel ;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;;
   2     siegel
   2     siegel ;Rules for sentence chains or single sentences
   2     siegel
   2     siegel ;declarative sentence, finite verb
   2     siegel
 334 francis_bond ; <ex>(do-parse-tty "食べる")</ex>
 334 francis_bond utterance_rule-decl-finite := utterance-sf-type &
 424 francis_bond  [SYNSEM.LOCAL.CAT.HEAD.SMOD decl,
 334 francis_bond   C-CONT.HOOK.INDEX.SF prop,
 334 francis_bond   ARGS.FIRST.SYNSEM [LOCAL.CAT.HEAD [MODUS uttmodus,
 334 francis_bond       FIN +],
 334 francis_bond       NON-LOCAL.QUE <! !>]].
```

It shows the revision, who committed it and which lines were affected.

Tracking Changes in MS Word



Collaborative Authoring Strategies

- Everyone writes a version and someone merges
- Pass a file round
- Have a copy at a shared repository (checkout/commit)
- Allow simultaneous (nearly) editing online
  only made possible by fast reliable internet
  - the wiki: a website that allows the easy creation and editing of any number of interlinked web pages via a web browser using a simplified markup language or a WYSIWYG text editor
  - Ward Cunningham, the developer of the first wiki software, WikiWikiWeb, originally described it as "the simplest online database that could possibly work."
  - "Wiki" is a Hawaiian word for "fast"

Does collaborative authoring change language?

- No one has measured this (as far as I know)
  - Are sentences longer/shorter?
  - Is the writing easier/harder to follow?
  - Is it more/less pleasant to read?
  - Does it have more/fewer errors?

  Good topic for an undergraduate thesis, ...

- Collaboration is sometimes in large chunks, sometimes in small chunks, sometimes at the level of formatting

- This is an untapped area of study

- For computer science, the ability to make changes, share and revert easily has made a big difference to how work is done — I think this will also change at least some authorship (my research papers are written like this)

Wikis and Wikipedia

The original wiki

Wikipedia

- Nupedia
- Lowering the barrier to entry
- Who edits what?
- Who edits what when?
- How to ensure quality?
- Research with Wikipedia

Nupedia

- Build a free encyclopedia written by experts
  true experts in their fields who with few exceptions possess PhDs

- Nupedia had a seven-step editorial process, consisting of:

  1. Assignment
  2. Finding a lead reviewer
  3. Lead review
  4. Open review
  5. Lead copyediting
  6. Open copyediting
  7. Final approval and markup

- Wikipedia was a side-project to allow collaboration on articles prior to entering the peer review process (3.5)

- Nupedia never got beyond 24 articles

Lowering the barrier to entry

- Wikipedia makes it easy to share your knowledge
  - People like to do this
- You don't even have to register
- Getting content is hard — so it is crucial that it is easy to add information

Wikipedia has been an enormous success

- In a hysterical world, Wikipedia is a ray of light –and that's the truth: It has been the butt of jokes for years, but the online encyclopedia represents mankind at its very best (John Naughton, The Guardian 2020)

- Wikipedia edits have massive impact on tourism, say economists: Adding a few paragraphs and photos can boost revenue by £100,000 for small cities (The Guardian, 2020)

- Wikipedia is open content, released under a free license. Knowing this encourages people to contribute; they know it's a public project everyone can use.

- Wikipedia's neutral point of view policy makes it an excellent place to gain a quick understanding of controversial topics.

- Articles steadily become more polished as they develop, ...

Quoted directly from Wikipedia:Why Wikipedia is so great unless otherwise noted.

Who writes Wikipedia: the in-crowd

- 50% of all the edits are done by just .7% of the users (524 people)
- 73% of all the edits are done by 2% (1,400 people)

Jimmy Wales (Talk at Stanford, 2005)

Most edits are done by insiders!

- Who Writes Wikipedia?
- Meet the most prolific contributor to the English version of wikipedia Washington Post (2019)

Who writes Wikipedia: the out-crowd

A close look at one page (**Alan Alda**):

If you just count edits (7 of the top 10) are registered users who (all but 2) have made thousands of edits to the site. Indeed, #4 has made over 7,000 edits while #7 has over 25,000.

When you count letters: few of the contributors (2 out of the top 10) are even registered and most (6 out of the top 10) have made less than 25 edits to the entire site. In fact, #9 has made exactly one edit —this one!

A great many people add a bit of content about something they know about, and a few people make many small changes to make it conform to wikipedia style.

Most content is added by outsiders!

Aaron Swartz (2006) Who Writes Wikipedia? *RAW THOUGHT* (weblog, accessed 2023-10-09)

Denise Anthony, Sean W Smith and Tim Williamson (2007) "The Quality of Open Source Production: Zealots and Good Samaritans in the Case of Wikipedia" (2007). Computer Science Technical Report TR2007-606.

WIERD Male Biases

- Wikipedia biases: Research exposes the male-dominated, pro-western worldview of the online encyclopedia Poppy Noor 2018-07-27
- Mainly Western
- Sources mainly in the language of the wikipedia
- Liberal bias
- History has a massive gender bias. We'll settle for fixing Wikipedia Monica Hesse 2019-02-17
- Only about 18 percent of Wikipedia's biographical articles are about women.
- That's up 3 percent from a few years ago,
- "contributors are majority Western and mostly male, and these gatekeepers apply their own judgment and prejudices" (Wikipedia Foundation)

The five pillars of Wikipedia

1. Wikipedia is an online encyclopedia
2. Wikipedia has a neutral point of view.
3. Wikipedia is free content
4. Wikipedians should interact in a respectful and civil manner
5. Wikipedia does not have firm rules

> the core aim of the Wikimedia Foundation, is to get a free encyclopedia to every single person on the planet.
>
> Jimmy Wales TED talk 2006

Encyclopedia

Wikipedia combines many features of general and specialized encyclopedias, almanacs, and gazetteers. Wikipedia is not a soapbox, an advertising platform, a vanity press, an experiment in anarchy or democracy, an indiscriminate collection of information, or a web directory. It is not a dictionary, a newspaper, or a collection of source documents, although some of its fellow Wikimedia projects are.

NPOV

We strive for articles in an impartial tone that document and explain major points of view, giving due weight for their prominence. We avoid advocacy, and we characterize information and issues rather than debate them. In some areas there may be just one well-recognized point of view; in others, we describe multiple points of view, presenting each accurately and in context rather than as "the truth" or "the best view". All articles must strive for verifiable accuracy, citing reliable, authoritative sources, especially when the topic is controversial or is about a living person. Editors' personal experiences, interpretations, or opinions do not belong on Wikipedia.

Free

Since all editors freely license their work to the public, no editor owns an article and any contributions can and may be mercilessly edited and redistributed. Respect copyright laws, and never plagiarize from any sources. Borrowing non-free media is sometimes allowed as fair use, but strive to find free alternatives first.

➢ licensed under the *Creative Commons Attribution-ShareAlike License*

Code of conduct and etiquette

Respect your fellow Wikipedians, even when you disagree. Apply Wikipedia etiquette, and do not engage in personal attacks. Seek consensus, avoid edit wars, and never disrupt Wikipedia to illustrate a point. Act in good faith, and assume good faith on the part of others. Be open and welcoming to newcomers. Should conflicts arise, discuss them calmly on the appropriate talk pages, follow dispute resolution procedures, and consider that there are over 6,726,000 other articles on the English Wikipedia to improve and discuss.

Ignore all rules (IAR)

Wikipedia has policies and guidelines, but they are not carved in stone; their content and interpretation can evolve over time. The principles and spirit matter more than literal wording, and sometimes improving Wikipedia requires making exceptions. Be bold, but not reckless, in updating articles. And do not agonize over making mistakes: (almost) every past version of a page is saved, so mistakes can be easily corrected.

Quality Control

- How is quality maintained?
- Some people regularly check the new changes page
- Many people watch pages of interest to them
- All information is stored and accessible
- Wikiscanner (an outside project links editors to the real world)
  - On November 17th, 2005, an anonymous Wikipedia user deleted 15 critical paragraphs from an article on e-voting machine-vendor Diebold
  - The editor was anonymous but the changes came from an IP address reserved for the corporate offices of Diebold

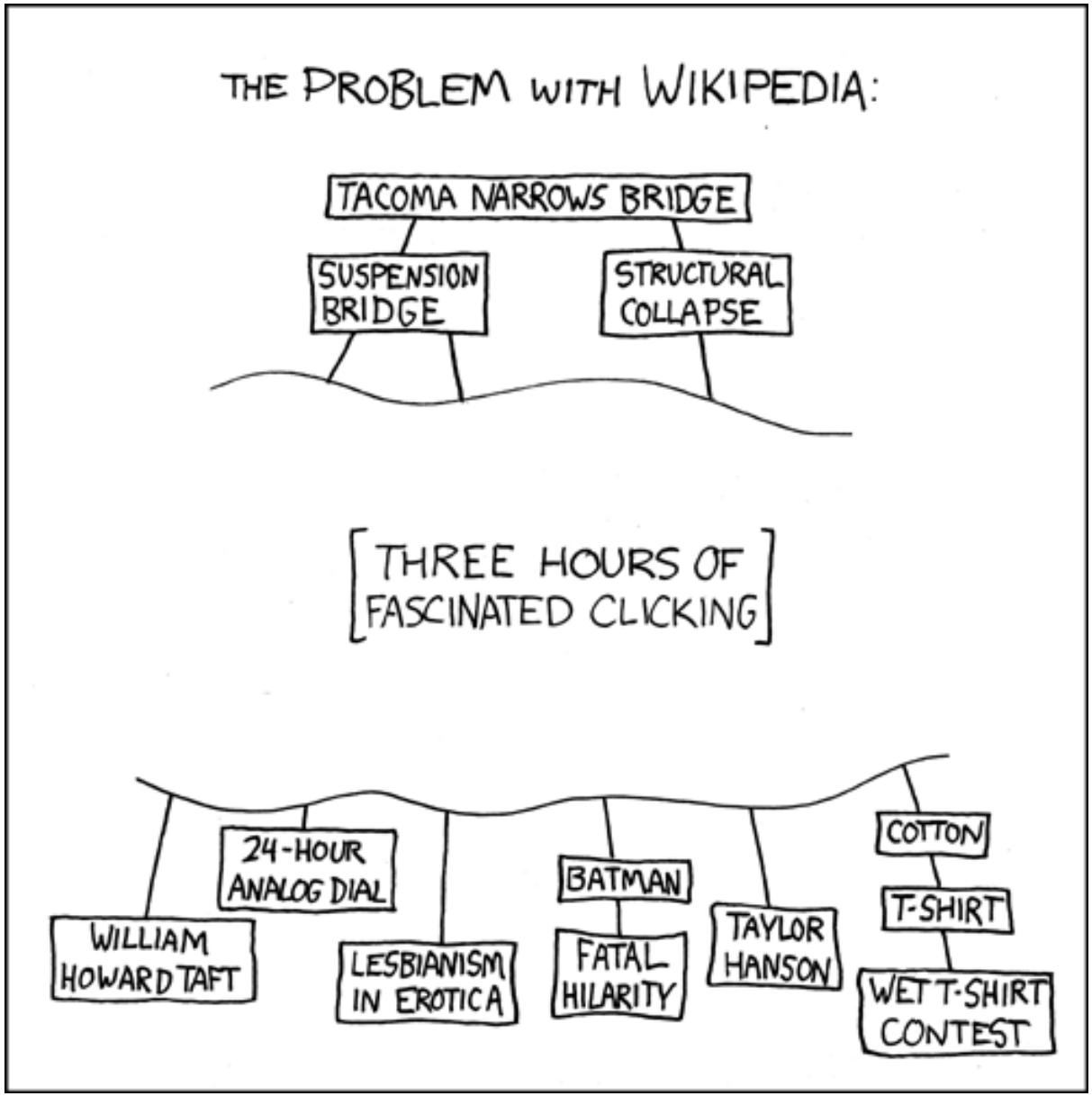
Wikipedia vs Britannica

- *Nature* study compared 42 articles reviewed by three experts
- The average scientific entry in Wikipedia contained four errors or omissions, while Britannica had three.
- Of eight "serious errors" the reviewers found —including misinterpretations of important concepts —four came from each source
- Wikipedia is good on pop culture and contemporary technology because Wikipedia's stable of dedicated volunteers tend to have more collective expertise in such areas.
- Wikipedia tends to lag when it comes to topics touching on the humanities
  - Techies are on-line more, and share information more readily.

Problems with Wikipedia

- Anyone can change an article in Wikipedia. Some articles in Wikipedia may not be entirely true and accurate, instead displaying a hoax or a false information.
- In particular, there is a problem of vandalism. Some is obvious, other forms may be difficult to see.
- People who have a strong opinion about a subject will try to control the articles about that subject.
- Not all facts are backed by sources, not all sources are reliable, and not all are checked.
- Not all editors are competent or pleasant.

Problems with Wikipedia (II)



Problems with Wikipedia (III)

The motto of Wikipedia should be

"the encyclopedia that anyone who understands the norms, socializes him or herself, dodges the impersonal wall of semi-automated rejection and still wants to voluntarily contribute his or her time and energy can edit."

Research with Wikipedia

- Extraction
  - Mining Wikipedia for structured knowledge
- Translation
  - Mining cross-wiki links for translations
- WeScience
  - Using Wikipedia as a vast corpus
- Training Data for Large Language Models

Other Wikis

- DELPH-IN wiki (Deep Linguistic Processing for HPSG Initiative)
- Wiktionary
- the Wiki of the Association for Computational Linguistics
- Conservapedia

...

What is a good article?

1. Well-written
2. Factually accurate and verifiable
3. Broad in its coverage
4. Neutral
5. Stable
6. Illustrated, if possible, by images

Well-written

- the prose is clear and concise, and the spelling and grammar are correct
  - Some students have been a bit sloppy about this in the past
- it complies with the manual of style guidelines for
  - lede/lead sections
  - layout
  - words to watch
  - fiction
  - list incorporation

Factually accurate and verifiable

- it provides references to all sources of information in the section(s) dedicated to the attribution of these sources according to the guide to layout;
- it provides in-line citations from reliable sources for direct quotations, statistics, published opinion, counter-intuitive or controversial statements that are challenged or likely to be challenged, and contentious material relating to living persons — science-based articles should follow the scientific citation guidelines;
- it contains no original research.
  - different from a research paper

Broad in its coverage

- it addresses the main aspects of the topic
- it stays focused on the topic without going into unnecessary detail

Neutral, Stable and Illustrated

- it represents viewpoints fairly and without bias
- it does not change significantly from day to day because of an ongoing edit war or content dispute
- images are tagged with their copyright status, and valid fair use rationales are provided for non-free content
- images are relevant to the topic, and have suitable captions.

Edit Wars

An edit war occurs when editors who disagree about some aspect of the content of a page repeatedly override each other's contributions, rather than try to resolve the disagreement by discussion.

Edit warring is unconstructive and creates animosity between editors, making it harder to reach a consensus as to the right way to improve the encyclopedia. Users who engage in edit wars risk being blocked or even banned from editing.

The three-revert rule (3RR): do not perform more than three reverts on a single page within a 24-hour period. Breaking this rule is sufficient—but not necessary—to warrant a block for edit warring.

Lamest Edit Wars

German Wikipedia's Article on Danube Tower Is the Danube Tower "an observation tower" or "a television and observation tower"? This edit war was so lame that it got covered in Der Spiegel.

Template:WikiProject Computer science 58kb of talk page debate plus a user block over how to copyedit a two line statement.

Sea of Japan Should it be called the Sea of Japan, the East Sea, or even the East Sea of Korea? Are both names valid, and if so, should the article be named Sea of Japan (East Sea) or Sea of Japan / East Sea? Or is the actual most common English and international name Sea of Japan (East Sea), parentheses and all? Should the dispute page be called the Sea of Japan naming dispute, or the Naming dispute over the body of water between Japan and Korea and the Russian Far East? ...

Access and Ownership

Licenses and Ownership

- Copyright
- Copyleft
- Creative Commons
- Ownership of a Language

Copyright

- State assigned monopoly to encourage an author so that the author of a work may reap the fruits of his or her intellectual creativity for a limited period of time.
- Copyright is a form of protection provided by the laws of a country for original works of authorship, including literary, dramatic, musical, architectural, cartographic, choreographic, pantomimic, pictorial, graphic, sculptural, and audiovisual creations.
- First legislated in England (1710). There was no automatic copyright protection for unpublished works.
- Currently 70 years after death of author (varies from country to country)
- What is the best balance?

Copyleft

- the practice of using copyright law to offer the right to distribute copies and modified versions of a work and requiring that the same rights be preserved in modified versions of the work.
  - You can redistribute it if and only if others can also
- copyleft is a general method for making a program (or other work) free, and requiring all modified and extended versions of the program to be free as well
- Gnu General Public License (GPL)
  (Richard Stallman/Free Software Foundation)

Creative Commons (CC): some rights reserved

Attribution (BY) requiring attribution to the original author

Share Alike (SA) allowing derivative works under the same or a similar license (later version or different jurisdiction) (copyleft)

Non-Commercial (NC) requiring the work not be used for commercial purposes

No Derivative Works (ND) allowing only the original work, without derivatives

- Wikipedia is **CC BY SA**
- These slides are **CC BY**

Open Science

- the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional
  - publishing open research
  - making data available
  - making tools (mainly software available)
- NTU's policy: "The final research data from projects carried out at NTU shall be made available for sharing (via the NTU Data Repository) unless there are prior formal agreements with external collaborators and parties on non-disclosure or proprietary use of the data." NTU's default license is CC-BY-NC
- What does this mean for linguists?

Who owns a language?

Q : Why do depositors restrict access to material?

A : There are a variety of reasons why a depositor might specify an access restriction. The people they have recorded may specify that they want their material kept safe, but not distributed. Occasionally depositors restrict access for a limited time until they have checked the material, to see if it is accurate enough for distribution or publication —authors are concerned that inaccurate information may take on a life of its own. Some information (though only a small fraction of ASEDA holdings) is restricted because it is secret/sacred.

Q : Why do speakers restrict access to material in their languages?

A : Many speakers of endangered languages consider that their language is their intellectual property, passed down to them from their ancestors. If it is made freely available to others, then their rights in that language can be diminished. Usually they do not want strangers to use words and sentences of their languages in an inappropriate way, and want to be consulted prior to public use.

Conclusions

- Version Control Systems
  - Allow asynchronous cooperation, blur ownership
- Wikipedia
- Some issues within Wikipedia
- Licensing and Ownership

Project

- Now let's edit