

Halfak's Wiki Research Libraries

v 0.0.1



Outline

1. Openness: Data, algorithms & results
2. Current Libraries
 - a. `pip install mediawiki-utilities`
 - b. `pip install wikiclass`
 - c. `pip install deltas`
 - d. `pip install mwoauth`

Part 1

What is open research?

Three major artifacts

Three major artifacts

1. The manuscript: “Open Access”

ABS

American Behavioral Scientist

Home

OnlineFirst

All Issues

Subscribe

RSS 

Email Alerts

Search this journal

Advanced Journal Search »

Impact Factor: 0.622 | **Ranking:** 46/92 in Social Sciences, Interdisciplinary | 94/114 in Psychology, Clinical | **5-Year Ranking:** 34/92 in Social Sciences, Interdisciplinary | 77/114 in Psychology, Clinical

Source: 2012 Journal Citation Reports® (Thomson Reuters, 2013)

A [more recent](#) version of this article was published on [04-08-2013]



The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline

Aaron Halfaker aaron.halfaker@gmail.com

R. Stuart Geiger

Jonathan T. Morgan

John Riedl


Abstract

This Article

Published online before print
December 28, 2012, doi:
10.1177/0002764212469365

American Behavioral Scientist
December 28, 2012
0002764212469365

» **Abstract Free**

Full Text (PDF) 

All Versions of this Article:
Version of Record - Apr 8, 2013
» OnlineFirst Version of Record - Dec 28, 2012

What's this?

Current Issue

▶ August 2014, 58 (9)



▶ Alert me to new issues of
American Behavioral
Scientist



American Behavioral Scientist

Home

OnlineFirst

All Issues

Subscribe

RSS

Email Alerts

Impact Factor: 0.622 | Ranking: 46/92 in Social Sciences, Interdisciplinary | 94/114 in Psychology, Clinical | 5-Year Ranking: 34/92 in Social Sciences, Interdisciplinary | 77/114 in Psychology, Clinical

Source: 2012 Journal Citation Reports® (Thomson Reuters, 2013)

A more recent version of this article was published on [04-08-2013]



The Rise and Decline of an Open Collaboration System: How Wikipedia's Reaction to Popularity Is Causing Its Decline

Aaron Halfaker aaron.halfaker@gmail.com

R. Stuart Geiger

Jonathan T. Morgan

John Riedl

Abstract

This Article

Published online before print
December 28, 2012, doi:
10.1177/0002764212469365

American Behavioral Scientist
December 28, 2012
0002764212469365

» Abstract Free

Full Text (PDF)



All Versions of this Article:
Version of Record - Apr 8, 2013
» OnlineFirst Version of Record - Dec 28, 2012

What's this?

Current Issue

» August 2014, 58 (9)



» Alert me to new issues of American Behavioral Scientist

To view this item, select one of the options below:

› Sign In

Already an individual subscriber?

If so, please sign in to American Behavioral Scientist with your User Name and Password.

User Name

Sign In

Password

Remember my user name & password.

[Forgot your user name or password?](#)

› [Can't get past this page?](#)

› [Help with Cookies.](#)

› [Need to Activate?](#)

› Purchase Short-Term Access

- › [Pay per Article](#) - You may purchase **this article** for US\$30.00. You must download your purchase, which is yours to keep, within 24 hours.
- › [Regain Access](#) - You can regain access to a recent Pay per Article purchase if your access period has not yet expired.

› OpenAthens Users

- › [Sign in via OpenAthens](#) : If your organization uses OpenAthens, you can log in using your OpenAthens username and password. Contact your library for more details.
- › [List of OpenAthens registered sites](#), including contact details.

› Login via Your Institution

- › [Login via your institution](#) : You may be able to gain access using your login credentials for your institution. Contact your library if you do not have a username and password.

› Subscribe/Recommend

- › [Click here](#) to subscribe to the print and/or online journal.
- › [Click here](#) to recommend to your library.

DEC 20, 2012

What's this?

Services

- › [Email this article to a colleague](#)
- › [Alert me when this article is cited](#)
- › [Alert me if a correction is posted](#)
- › [Similar articles in this journal](#)
- › [Download to citation manager](#)
- › [Request Permissions](#)
- › [Request Reprints](#)
- › [Load patientINFORMATION](#)

Citing Articles

- › [Load citing article information](#)
- › [Citing articles via Scopus](#)
- › [Citing articles via Google Scholar](#)

Google Scholar

- › [Articles by Halfaker, A.](#)
- › [Articles by Riedl, J.](#)
- › [Search for related content](#)

Related Content

Load related web page information

- › [Journal Home](#)
- › [Subscriptions](#)
- › [Archive](#)
- › [Contact Us](#)
- › [Table of Contents](#)

To view this item, select one of the options below:

➤ Sign In

Already an individual subscriber?

If so, please sign in to American Behavioral Scientist with your User Name and Password.

User Name

Sign In

Password

Remember my user name & password.

[Forgot your user name or password?](#)

➤ [Can't get past this page?](#)

➤ [Help with Cookies.](#)

➤ [Need to Activate?](#)

Dec 20, 2012

What's this?

Services

- [Email this article to a colleague](#)
- [Alert me when this article is cited](#)
- [Alert me if a correction is posted](#)
- [Similar articles in this journal](#)
- [Download to citation manager](#)
- [Request Permissions](#)
- [Request Reprints](#)
- [Load patientINFORMATION](#)

Citing Articles

You may purchase this article for US\$30.00.

➤ OpenAthens Users

- [Sign in via OpenAthens](#) : If your organization uses OpenAthens, you can log in using your OpenAthens username and password. Contact your library for more details.
- [List of OpenAthens registered sites](#), including contact details.

➤ Login via Your Institution

- [Login via your institution](#) : You may be able to gain access using your login credentials for your institution. Contact your library if you do not have a username and password.

➤ Subscribe/Recommend

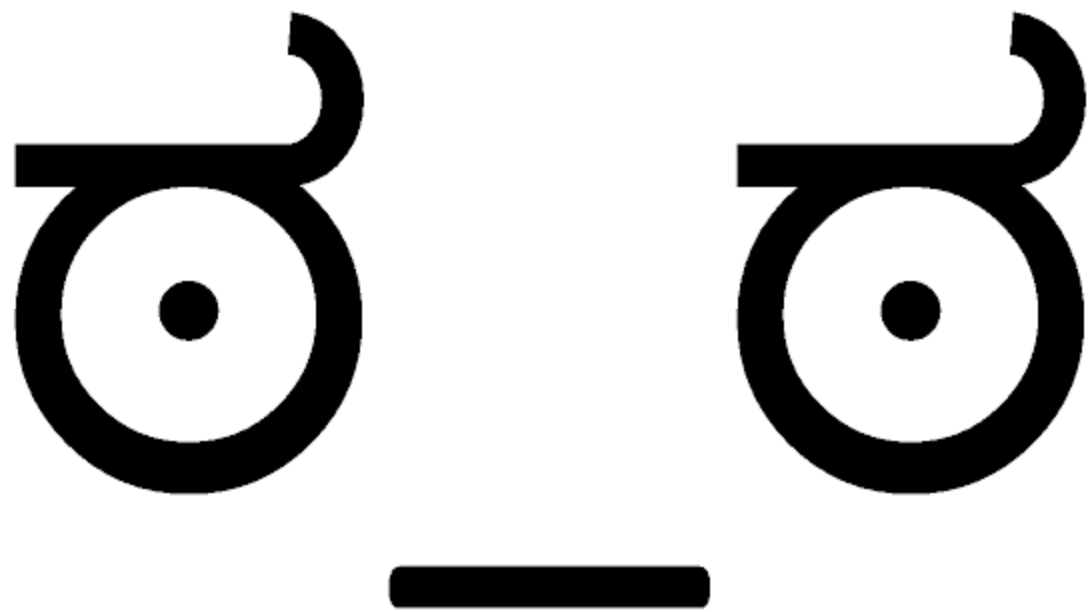
- [Click here](#) to subscribe to the print and/or online journal.
- [Click here](#) to recommend to your library.

➤ [Search for related content](#)

Related Content

Load related web page information

- [Journal Home](#)
- [Subscriptions](#)
- [Archive](#)
- [Contact Us](#)
- [Table of Contents](#)



My hack

Open access summary

(I am not a lawyer)

Summary

Gentle reader,

Below is the authors' summary of a paper that was accepted to a special issue of [American Behavioral Scientist](#) on Wikis. I also provide the [unabridged, pre-print of the paper](#) if you desire more discussion and explanation.

Feel free to email me with questions, and please let me know if you find an error.

Enjoy!

Aaron Halfaker (aaron.halfaker@gmail.com)

According to [a report published in 2009](#) by the Wikimedia Foundation, the number of active editors working on the English Wikipedia is declining. As the figure 1 below suggests, the number of active editors (editors with ≥ 5 edits/month) abruptly stopped growing in early 2007 and entered a steady, linear decline. Recent research has shown evidence that this transition is rooted in the declining retention of new editors, not a change in the retention of already-experienced old-timers (Suh, 2009). What is unclear, or was before this work, is why this sudden change in the retention of new editors took place.

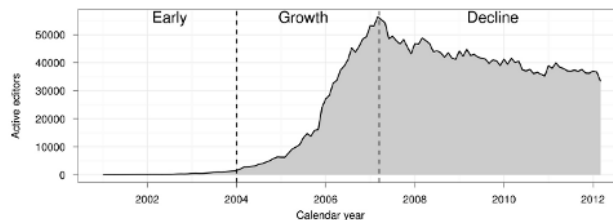


Figure 1. The editor decline. The number of active editors (≥ 5 edits/month) is plotted over time for the English language Wikipedia.

This paper implicates the strategies adopted by Wikipedia editors to preserve the quality and consistency of the encyclopedia in causing the decline in retention of desirable newcomers. Below, the results are broken up into a description of three general findings.

The decline in desirable newcomers

One of the biggest open questions about Wikipedia's newcomer decline was whether it was the result of a natural decline in the quality of newcomers (where lower-quality newcomers were "encouraged" to go elsewhere) or whether changes in how Wikipedia welcomes newcomers was at fault. In order to explore this, we [manually categorized](#) the work of 2100 newcomers sampled over the history of the website. With the help of Maryana Pinchuk ([Accedie](#)), Oliver Keyes ([Ironholds](#)) and Steven Walling ([Steven Walling](#)) we categorized these newcomers into 4 ordinal quality classes based on their first session of editing activity:

My current process

Research

Discussion

Read

Edit

View history



More ▾

Search



Research:Wikipedia article creation

Code repository: <https://github.com/halfak/Wikipedia-article-creation-research>

The process of creating articles is becoming increasingly difficult for new users due to increasingly restrictive criteria^[1] and the speed at which their articles are tagged and deleted^[2]. This trend is concerning because new users tend to leave the wiki when their work is deleted.

The [English Wikipedia Articles for Creation WikiProject](#) has recently adjusted in order to encourage new editors to create draft articles outside of the usual article space. However, it's unclear whether such initiatives are successful in improving the success rate of articles created by new editors or improving their retention. In this study, we'll discuss our analysis of newcomer created articles in the most active Wikipedia projects and answer questions about how different workflows affect the success rate of articles.

Contents [\[show\]](#)

Related work [\[edit\]](#)

Research has established that the number of active editors in the English Wikipedia has entered a decline and that this decline is the result of decreased retention of new users^[3]. Subsequent research by Halfaker et al. has shown evidence that this decline is not due to the quality of newcomers, but rather the increasing complexity newcomers must manage in order to successfully contribute and the negative reactions they receive^[4]. One of the key factors in Halfaker et al.'s model predicting the retention of new editors was whether they created articles that were quickly deleted. Related work by [User:Mr.Z-man](#) confirmed that new editors who created articles that were deleted are less likely to continue to contribute^[5]. Research performed in parallel found that the rate at which newly created articles are deleted has risen sharply in recent years^[1] and the speed at which new articles are tagged and deleted has increased dramatically^[2].

Research project

Article creation

Main contact **Halfak (WMF)**
Wikimedia Foundation

Co-investigators Steven Walling

Start 2013-09

Status completed

Open access

WMF support **HO**

Wikimedia research projects

Three major artifacts

1. The manuscript: “Access”
2. The data: “Data”

Content page

[Discussion](#)

Read

[Edit](#)

[View history](#)

Search



Data dumps

Contents [\[show\]](#)

Summary [\[edit\]](#)

Description [\[edit\]](#)

WMF publishes data dumps of Wikipedia and all WMF projects on a regular basis. English Wikipedia is dumped once a month, while smaller projects are often dumped twice a month.

Content [\[edit\]](#)

- Text and metadata of current or all revisions of all pages as XML files
- Most database tables as sql files
 - Page-to-page link lists (pagelinks, categorylinks, imagelinks, templatelinks tables)
 - Lists of pages with links outside of the project (externallinks, iwlinks, langlinks tables)
 - Media metadata (image, oldimage tables)
 - Info about each page (page, page_props, page_restrictions tables)
 - Titles of all pages in the main namespace, i.e. all articles (*-all-titles-in-ns0.gz)
 - List of all pages that are redirects and their targets (redirect table)

[Main page](#)

[Wikimedia News](#)

[Translations](#)

[Recent changes](#)

[Random page](#)

[Help](#)

[Babel](#)

Community

[Wikimedia Forum](#)

[Mailing lists](#)

[Requests](#)

[Babylon](#)

[Reports](#)

[Research](#)

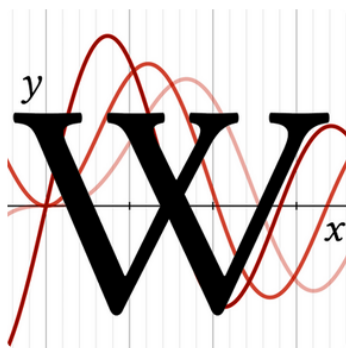
[Planet Wikimedia](#)

Beyond the Web

[Meet Wikimedians](#)

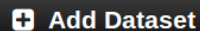
[Events](#)

[Chapters](#)



Wikimedia

A collection of datasets about Wikipedia and other projects run by the Wikimedia Foundation. The collection is open to contributions by researchers not affiliated with

 Datasets Activity Stream About

Search datasets...

**15 datasets found**

Order by: Name Ascending ▾

Teahouse corpus

The Teahouse corpus is a set of questions asked at the Wikipedia Teahouse, a peer support forum for new Wikipedia editors. This corpus contains data from its first two years of...

<http://datahub.io/dv/group/wikimedia>

See also

- [Teahouse project documentation](#): project planning docs and reports from the Teahouse pilot
- [Wikimedia data portal](#): public data resources on Wikipedia and other Wikimedia projects
- [Mediawiki database schema](#): description of standard data tables and fields in MediaWiki sites
- [Mediawiki API documentation](#): data available through the MediaWiki API (depending on site configuration)
- [Wikitext markup information](#): information about the markup conventions used in the text of the Teahouse corpus

Data and Resources



Teahouse questions - 2/23/2014

Metadata for 5,003 questions. See README: Teahouse questions for field...



Preview



Download



Teahouse question text - 2/23/2014

The raw text of 4,998 questions. See README: Teahouse question text for...



Preview



Download



README: Teahouse question text

Field types and value definitions for teahouse-question-text datafile



Preview



Download



README: Teahouse questions

Data field types and values



Preview



Download



[en>User:Jtmorgan](#)

Three major artifacts

1. The manuscript: “Access”
2. The data: “Data”
3. The code: “Source”

Why open source?

Red Hat® believes open source simply creates better software. Everyone collaborates.

The best technology wins. Not just within one company, but for everyone, anyone, around the world. <http://www.redhat.com/about/whoisredhat/opensource.html>

... but for science?

Why open source?

HALFAK believes open source simply creates better **SCIENCE**. Everyone collaborates.

The best **SCIENCE HAPPENS**. Not just within one company, but for everyone, anyone, around the world. --Halfak

... but for science?

Part 2

Libraries!

Libraries

- **MediaWiki Utilities** -- General data processing
 - Repo: <https://github.com/halfak/Mediawiki-Utilities>
 - Docs: <http://pythonhosted.org/mediawiki-utilities>
- **Wiki-Class** -- Article quality classification
 - Repo: <https://github.com/halfak/Wiki-Class>
 - Docs: <https://pythonhosted.org/wikiclass>
- **MediaWiki OAuth** -- OAuth handshaker
 - Repo: <https://github.com/halfak/MediaWiki-OAuth>
 - Docs: <http://pythonhosted.org/mwoauth>
- **Deltas** -- Robust difference detection
 - Repo: <https://github.com/halfak/Deltas>
 - Docs: <http://pythonhosted.org/deltas>

OAuth!

Allows Wiki-tool users to log into Wikipedia with their accounts.



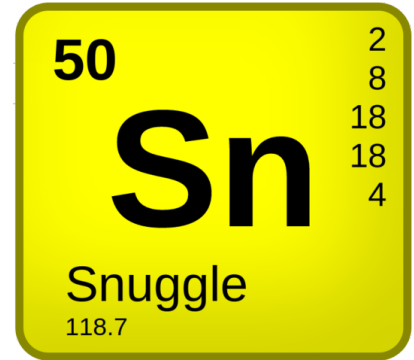
Wikimetrics

Welcome to the Wikimedia Foundation's Wikimetrics homepage. This API allows you to select a set of users, also known as a "cohort" (for example, all users who signed up via the Thank You campaign) select a metric to be computed for each of these users (for example, how many [bytes they've added](#)) with optional parameters (for example, a time range) and retrieve the response in JSON or CSV format.

You can also compute a single, aggregate value for the cohort (like the mean revert rate) .

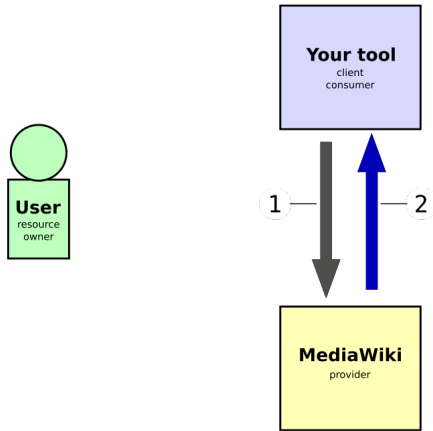
[Learn More](#)

[Analyze](#)

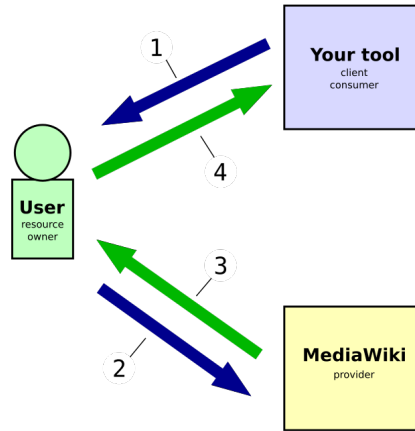


OAuth Handshake

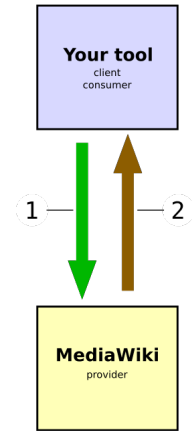
1. Initialize



2. Authorize

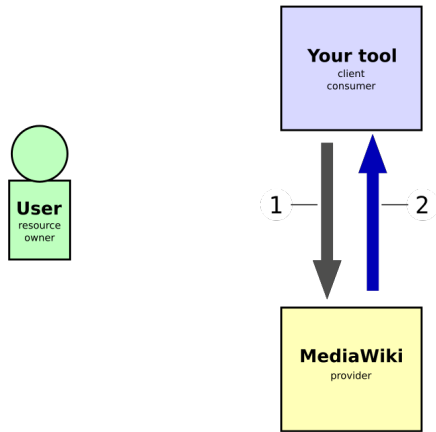


3. Complete



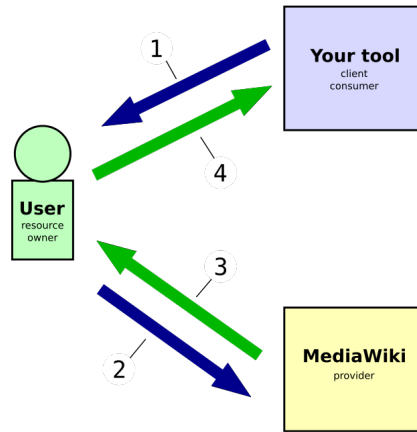
OAuth Handshake

1. Initialize



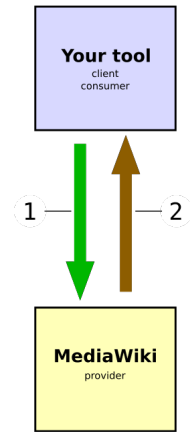
Request token

2. Authorize



Verifier

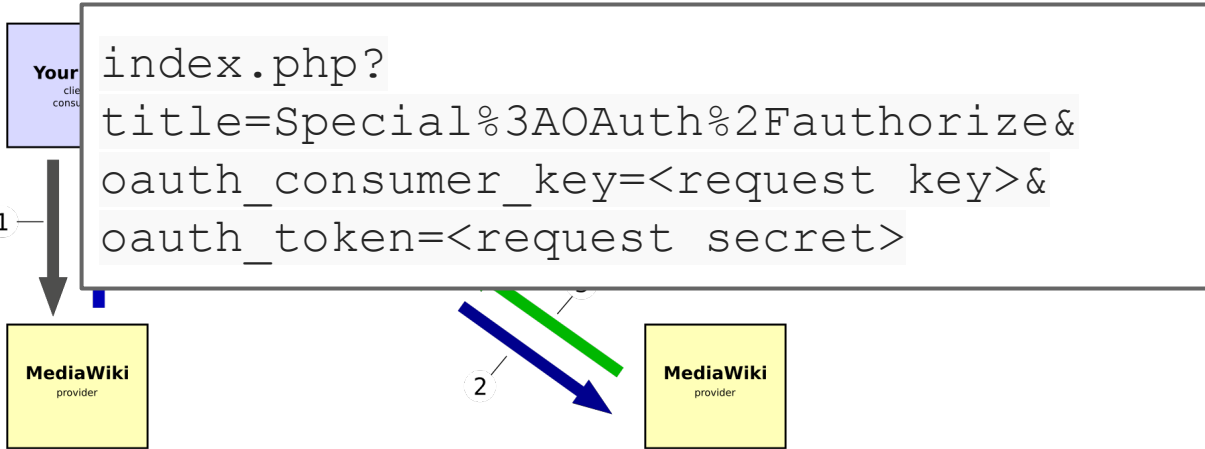
3. Complete



Access token

OAuth Handshake

1. Initialize

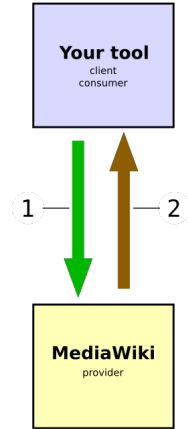


Request token

2. Authorize

Verifier

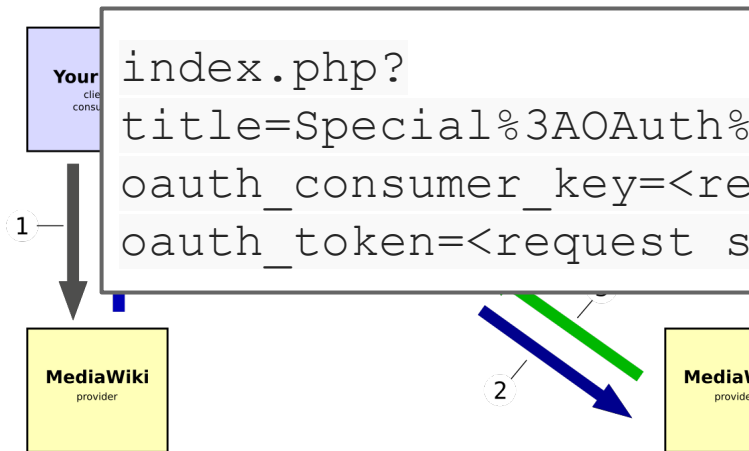
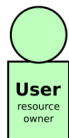
3. Complete



Access token

OAuth Handshake

1. Initialize



Request token

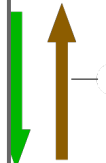
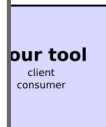
2. Authorize

Verifier

JSON Web Tokens

Table of Contents

1. Introduction
 - 1.1. Notational Conventions
 2. Terminology
 3. JSON Web Token (JWT) Overview
 - 3.1. Example JWT
 4. JWT Claims
 - 4.1. Registered Claim Names
 - 4.1.1. "iss" (Issuer) Claim
 - 4.1.2. "sub" (Subject) Claim
 - 4.1.3. "aud" (Audience) Claim
 - 4.1.4. "exp" (Expiration Time) Claim
 - 4.1.5. "nbf" (Not Before) Claim
 - 4.1.6. "iat" (Issued At) Claim
 - 4.1.7. "jti" (JWT ID) Claim
 - 4.2. Public Claim Names
 - 4.3. Private Claim Names
 5. JOSE Header
 - 5.1. "typ" (Type) Header Parameter
 - 5.2. "cty" (Content Type) Header Parameter
 - 5.3. Replicating Claims as Header Parameters
 6. Plaintext JWTs
 - 6.1. Example Plaintext JWT
 7. Rules for Creating and Validating a JWT
 - 7.1. String Comparison Rules
 8. Implementation Requirements
 9. URI for Declaring that Content is a JWT
 10. IANA Considerations
 - 10.1. JSON Web Token Claims Registry
 - 10.1.1. Registration Template
 - 10.1.2. Initial Registry Contents
 - 10.2. Sub-Namespace Registration of urn:ietf:params:oauth:token-type:jwt
 - 10.2.1. Registry Contents
 - 10.3. Media Type Registration
 - 10.3.1. Registry Contents
 - 10.4. Header Parameter Names Registration
 - 10.4.1. Registry Contents
 11. Security Considerations
 - 11.1. Trust Decisions
 - 11.2. Signing and Encryption Order
 12. Privacy Considerations
 13. References
 - 13.1. Normative References
 - 13.2. Informative References
- Appendix A. JWT Examples**
- A.1. Example Encrypted JWT
 - A.2. Example Nested JWT
- Appendix B. Relationship of JWTs to SAML Assertions**
- Appendix C. Relationship of JWTs to Simple Web Tokens (SWTs)**
- Appendix D. Acknowledgements**
- Appendix E. Document History**
- § Authors' Addresses**



1. I

Check the *issuer*

Confirm that token.iss matches the domain of MediaWiki that you made the request to (e.g. "mediawiki.org")

Check the *audience*

Confirm that token.aud matches your **consumer key**

Check the *issued at time*

Confirm that token.iat (unix timestamp in seconds) is before the current time

Check the *expiration time*

Confirm that token.exp (unix timestamp in seconds) is after the current time

Check the *number used only once*

Confirm that token.nonce matches the nonce your application set with the original request



User resource owner

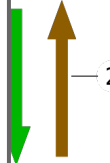
Re

JSON Web Tokens

Table of Contents

- 1. Introduction
 - 1.1. Notational Conventions
- 2. Terminology
- 3. JSON Web Token (JWT) Overview
 - 3.1. Example JWT
- 4. JWT Claims
 - 4.1. Registered Claim Names
 - 4.1.1. "iss" (Issuer) Claim
 - 4.1.2. "sub" (Subject) Claim
 - 4.1.3. "aud" (Audience) Claim
 - 4.1.4. "exp" (Expiration Time) Claim
 - 4.1.5. "nbf" (Not Before) Claim
 - 4.1.6. "iat" (Issued At) Claim
 - 4.1.7. "jti" (JWT ID) Claim
 - 4.2. Public Claim Names
 - 4.3. Private Claim Names
- 5. JOSE Header
 - 5.1. "typ" (Type) Header Parameter
 - 5.2. "cty" (Content Type) Header Parameter
 - 5.3. Replicating Claims as Header Parameters
- 6. Plaintext JWTs
 - 6.1. Example Plaintext JWT
- 7. Rules for Creating and Validating a JWT
 - 7.1. String Comparison Rules
- 8. Implementation Requirements
- 9. URI for Declaring that Content is a JWT
- 10. IANA Considerations
 - 10.1. JSON Web Token Claims Registry
 - 10.1.1. Registration Template
 - 10.1.2. Initial Registry Contents
 - 10.2. Sub-Namespace Registration of urn:ietf:params:oauth:token-type:jwt
 - 10.2.1. Registry Contents
 - 10.3. Media Type Registration
 - 10.3.1. Registry Contents
 - 10.4. Header Parameter Names Registration
 - 10.4.1. Registry Contents
- 11. Security Considerations
 - 11.1. Trust Decisions
 - 11.2. Signing and Encryption Order
- 12. Privacy Considerations
- 13. References
 - 13.1. Normative References
 - 13.2. Informative References
- Appendix A. JWT Examples
 - A.1. Example Encrypted JWT
 - A.2. Example Nested JWT
- Appendix B. Relationship of JWTs to SAML Assertions
- Appendix C. Relationship of JWTs to Simple Web Tokens (SWTs)
- Appendix D. Acknowledgements
- Appendix E. Document History
- § Authors' Addresses

our tool
client
consumer



MediaWiki
provider

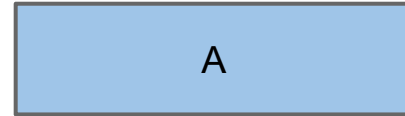
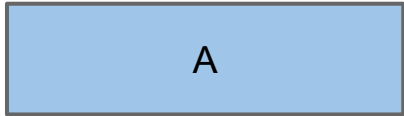
2



https://commons.wikimedia.org/wiki/File:Hair_pulling_stress.jpg

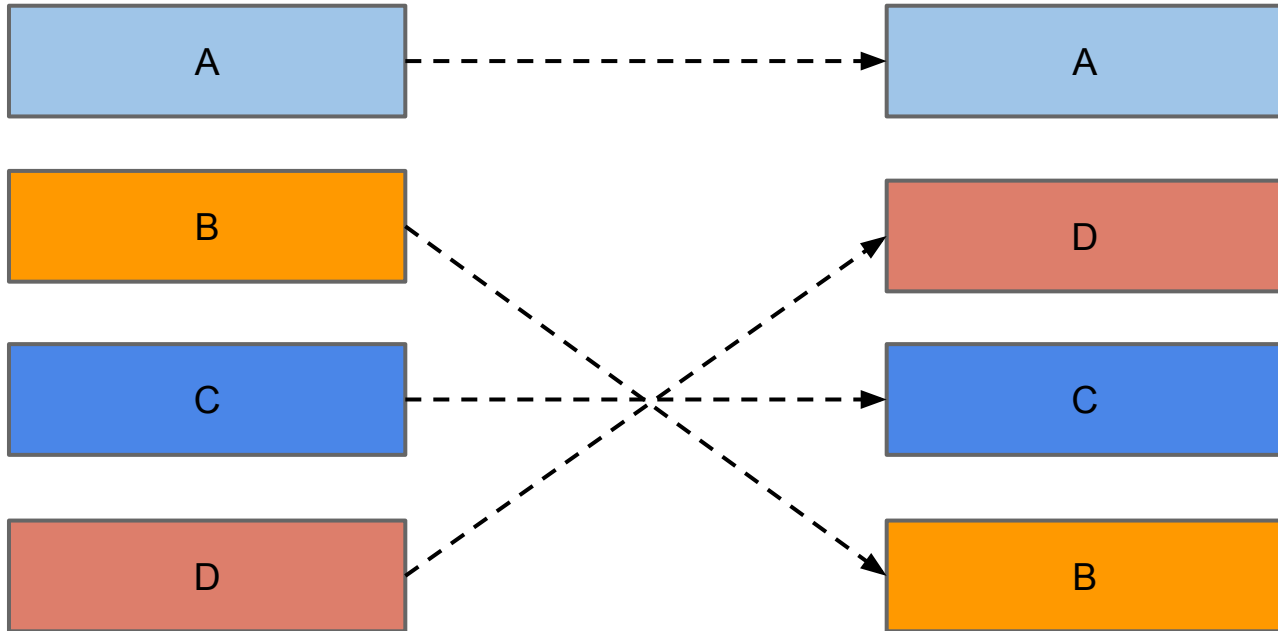
[back to the editor]

Difference algorithms!



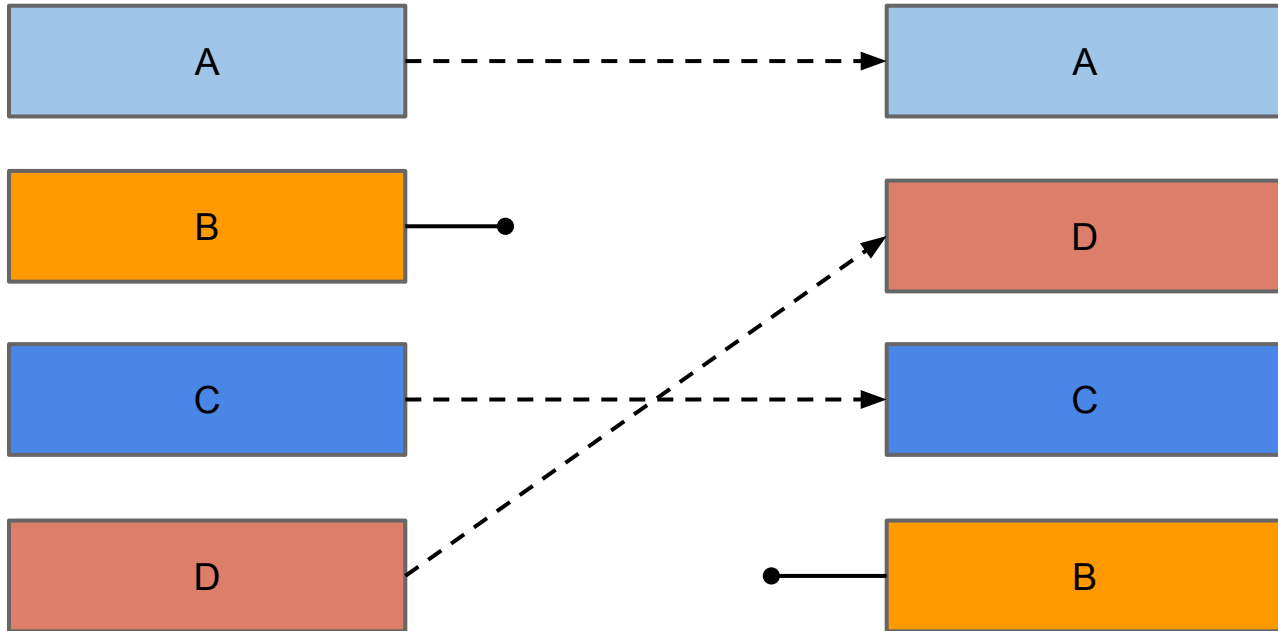
Difference algorithms!

Human intuition

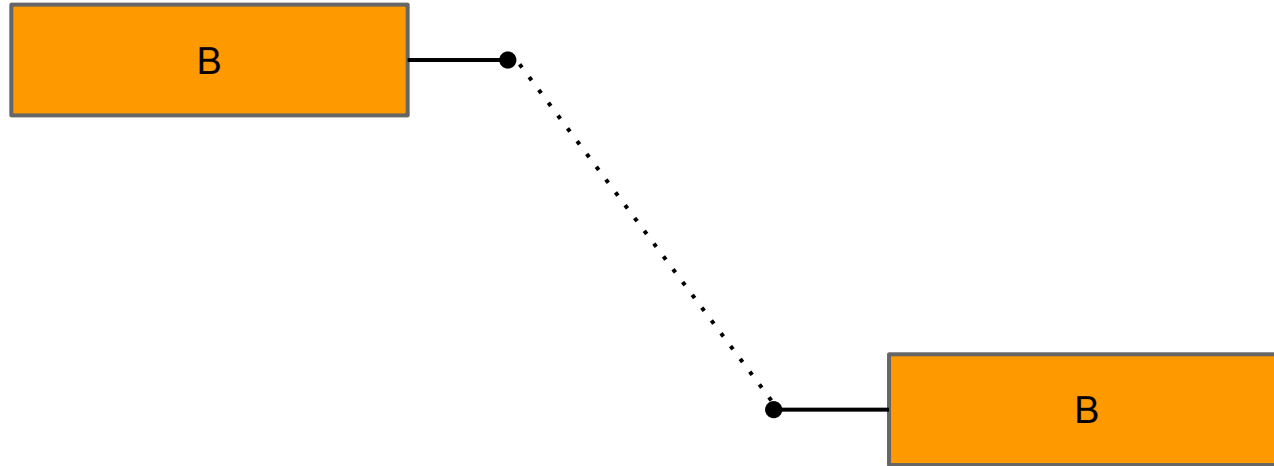


Difference algorithms!


Longest common substring



Attributing authorship of content?



Attributing authorship of content?



Dipl.-Medienwissenschaftler Fabian Flöck

Research Associate
Phone: +49 721 608 4 6584
Telefax: +49 721 608 46580
Email: floeck#YouKnowWhatSymbol#kit.edu

Research Group: [Knowledge Management](#)
Room: 222 (Building: 11.40)

Fabian Flöck, Maribel Acosta

WikiWho: Precise and Efficient Attribution of Authorship of Revisioned Content

Proceedings of the 23rd international conference on World Wide Web, ACM, April, 2014

I am a professor of [Computer Science](#) at the [University of California, Santa Cruz](#).
This is my official UCSC home page.

Contact Information

- Email: luca@ucsc.edu
- Google Apps (Drive, Chat, Hangouts, ...): luca@ucsc.edu
- [My UCSC Google+ Profile](#)

I also have a [personal home page](#), for topics unrelated from UCSC.
Here is my [curriculum vitae](#) and [resume](#).

Interests

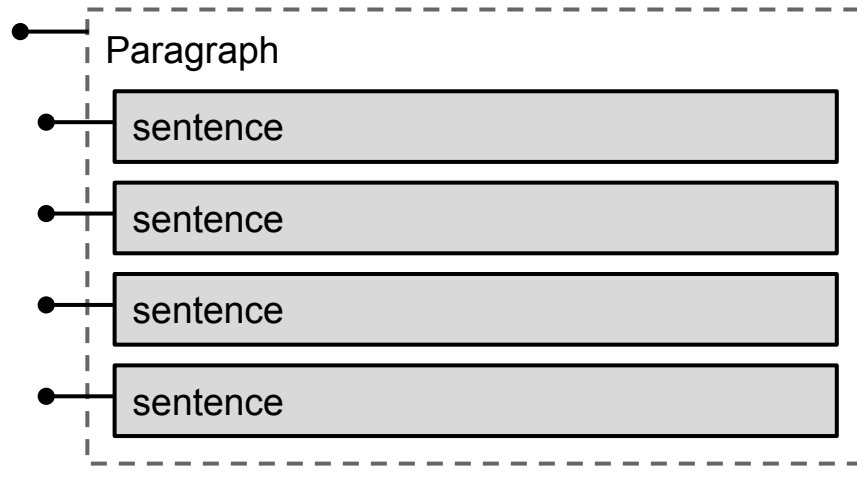
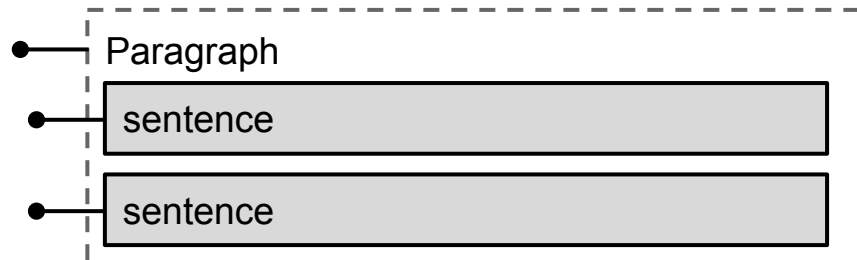
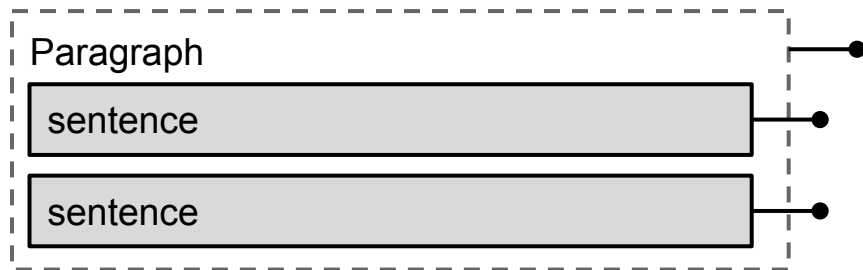
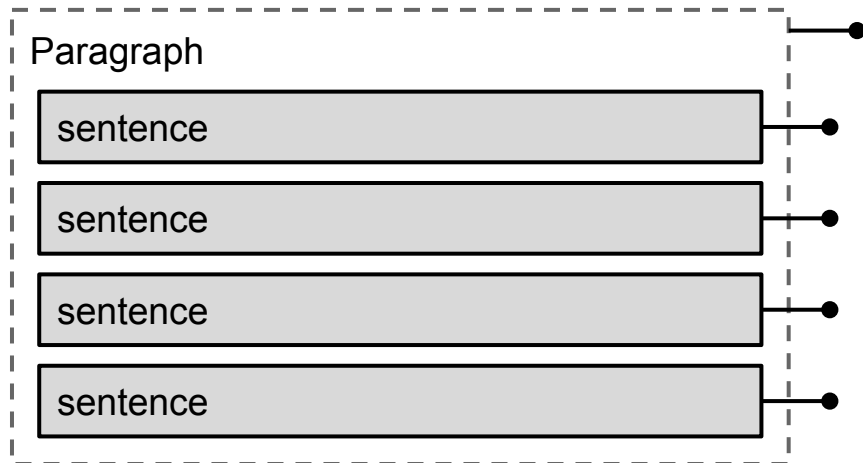


L. de Alfaro and M. Shavlovsky.

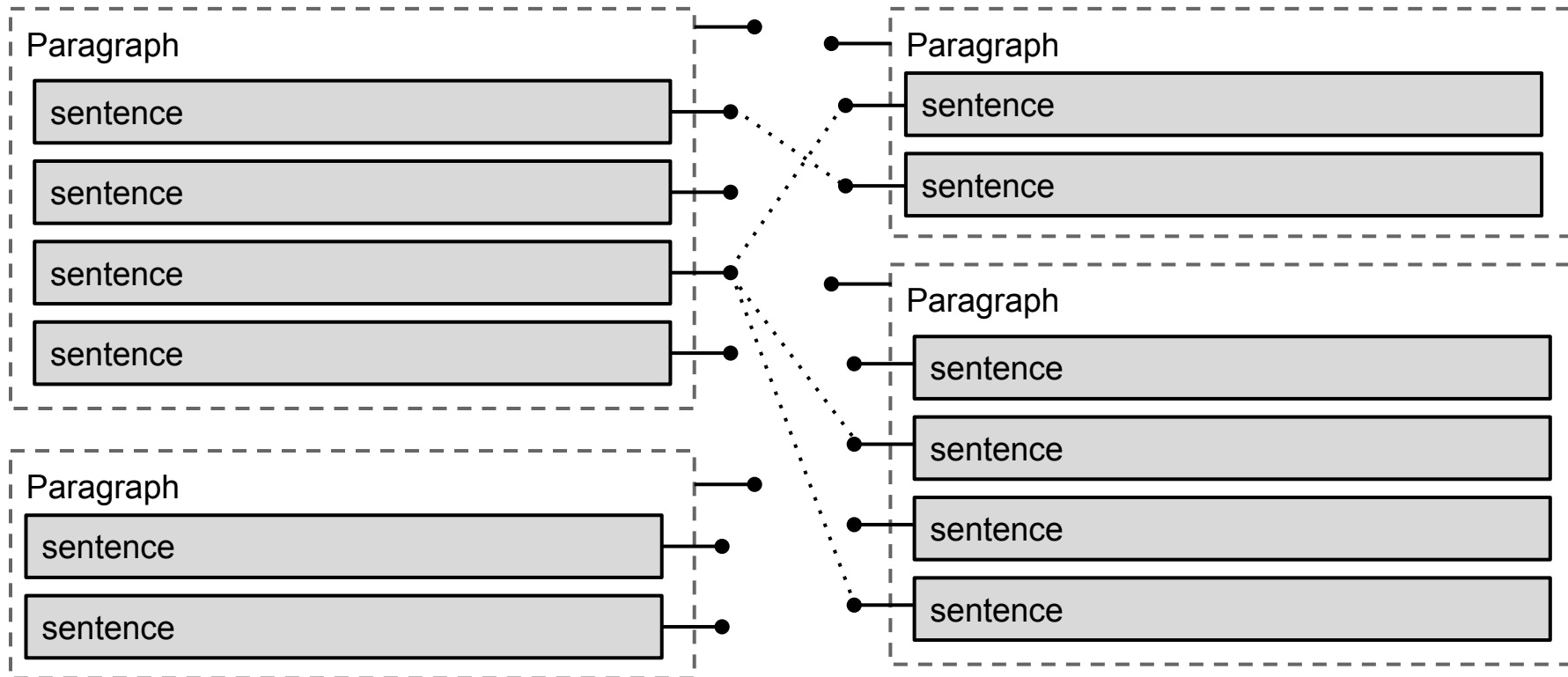
Attributing authorship of revisioned content.

Proceedings of the 22nd international conference on World Wide Web, ACM, April, 2013

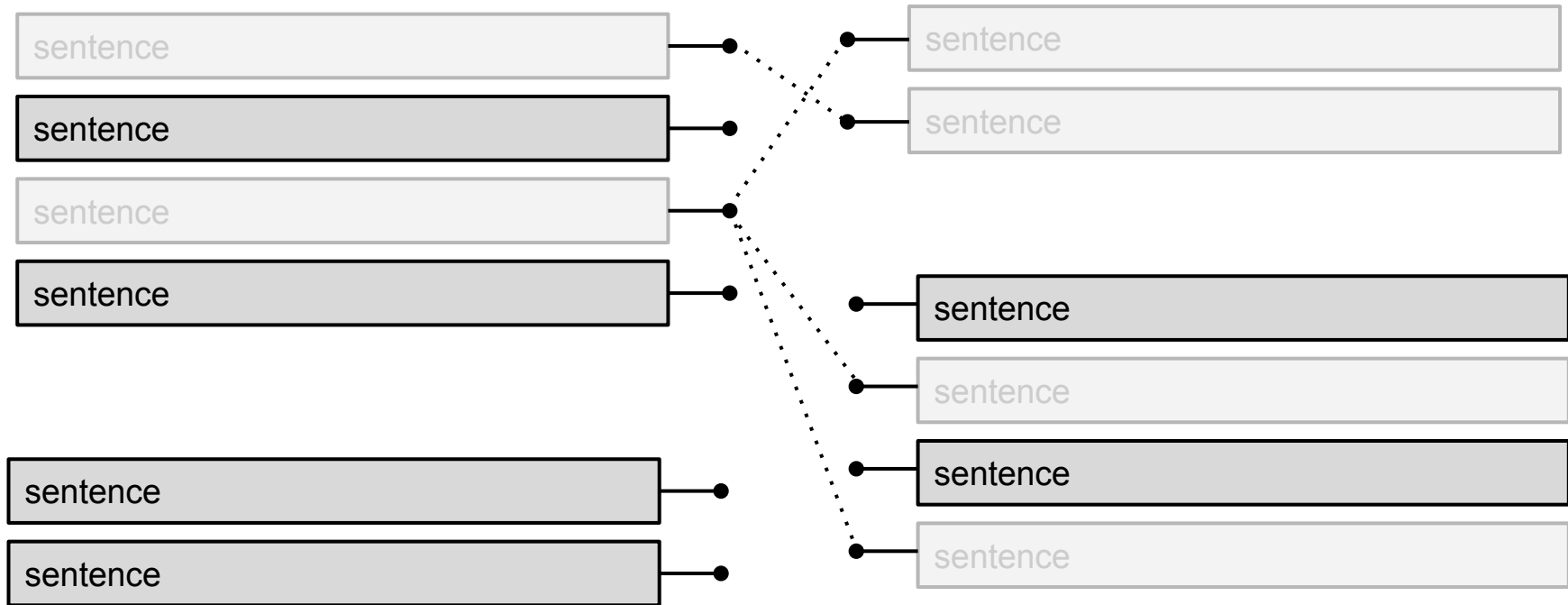
Segment matcher



Segment matcher

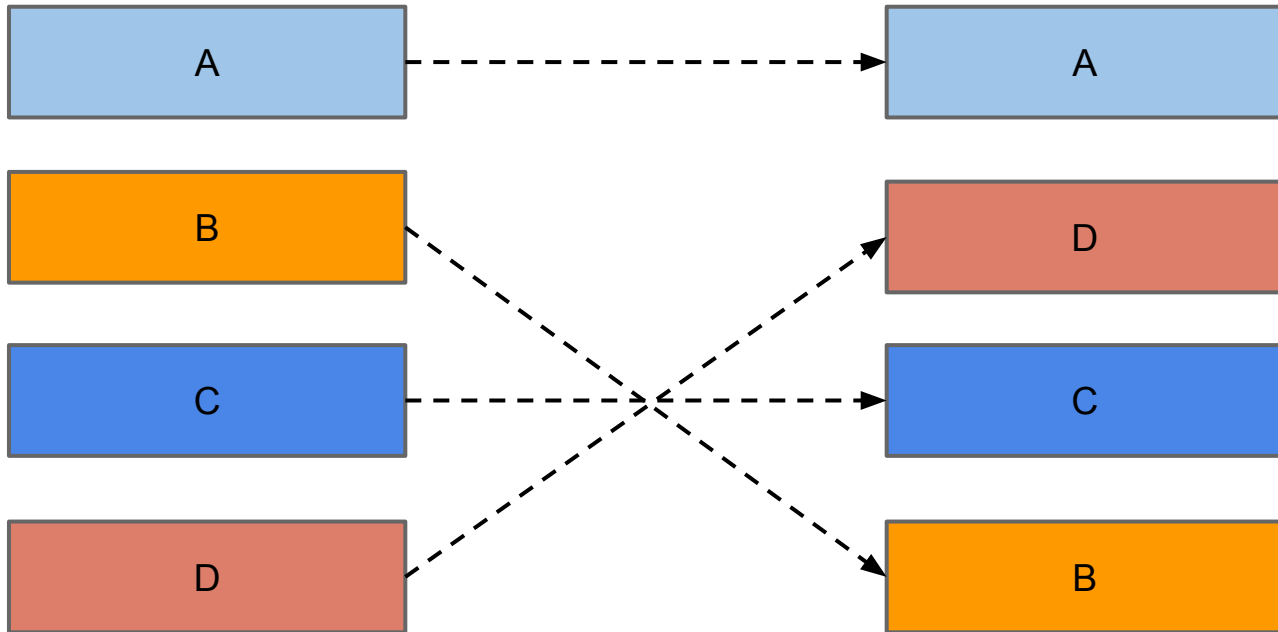


Segment matcher



Difference algorithms!

Segment Matcher == Human intuition



Research software as libraries

- Easy to re-use
 - `pip install mediawiki-utilities`

Research software as libraries

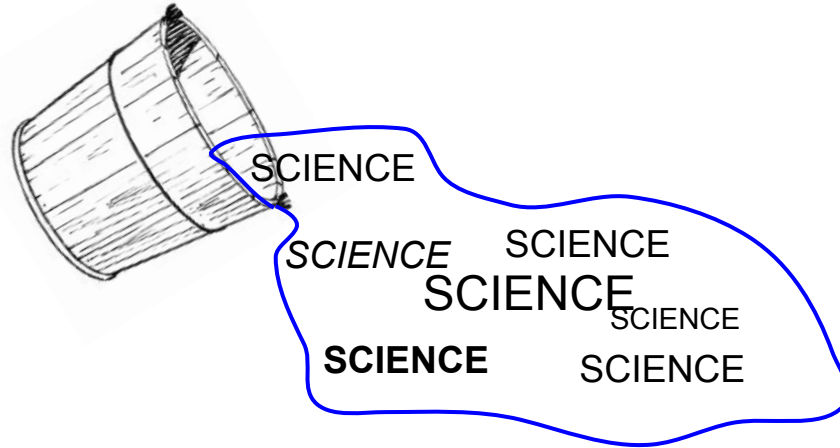
- Easy to re-use
 - `pip install mediawiki-utilities`
- More collaboration == more science
 - <http://github.com/halfak/mediawiki-utilities>
 - plz submit bugs and pull requests!

Research software as libraries

- Easy to re-use
 - `pip install mediawiki-utilities`
- More collaboration == more science
 - <http://github.com/halfak/mediawiki-utilities>
 - plz submit bugs and pull requests!
- Broad benefit & consistency
 - Complex problems
 - Bugs

Thanks!

Aaron Halfaker
ahalfaker@wikimedia.org
“halfak” on IRC
@halfak on Twitter



<http://github.com/halfak/Mediawiki-Utilities>
<http://github.com/halfak/Wiki-class>
<http://github.com/halfak/MediaWiki-OAuth>
<http://github.com/halfak/Deltas>

Props to:

- Yuvi Panda
- Filippo Valsorda
- Max Klien
- Morten Warncke-Wang
- Oliver Keyes

Send me pull requests!