Experimental Data for the A?B*A Pattern in CSS: Inputs and Outputs

Leonard Punt, Sjoerd Visscher, Vadim Zaytsev

@leonardpunt, @sjoerdvisscher, @grammarware

Where to find it

- **Description**
  - DOI [10.1109/ICSME.2016.91](https://doi.org/10.1109/ICSME.2016.91)
  - page 616
  - © IEEE

Experimental Data for the A?B*A Pattern in CSS: Inputs and Outputs

@leonardpunt
Leonard Punt
University of Amsterdam, The Netherlands
Q42, The Netherlands

@sjoerdvisscher
Sjoerd Visscher
Q42, The Netherlands

@grammarware
Vadim Zaytsev
University of Amsterdam, The Netherlands
Raincode, Belgium

The complete dataset is available under the MIT license from this web location: http://leonardpunt.github.io/masterproject/dataset-and-results.zip. This dataset is used to detect undoing style in CSS code. In total, this dataset contains 41 subjects. Each subject has its own folder, which contains the captured states, a `states.html` file, is used to load all captured states in one document, and a folder called `results`, which contains the detected undoing styles, the refactored style sheets and the detected semantic changes. The file `states.html` was used as an input for our detection tool [4].

Selection of subjects. In order to select representative real-world web applications, we used the empirical data used in the study conducted by Mazinanian et al. [2]. This data set includes 38 popular online web applications and is available online [1]. Besides the 38 subjects from the study of Mazinanian et al., two web applications developed by Q42 have been studied as well. The 41st subject is a new version of the subject 'Gmail' from in the original data set. The complete list of selected systems follows:

Facebook	Pinterest	YouTube	Reddit
Twitter	Tumblr.com	YahooMail	Wordpress.org
Outlook.com	Vimeo.com	Gmail	Igloo
Github	Phormer	Amazon.ca	BeckerElectric
Ebay	Equus	About.com	ProToolsExpress
Alibaba	UniqueVanities	Apple.com	ICSE12
BBC	EmployeeSolutions	CNN	SyncCreative
Craiglist	GlobalTVBC	Imgur	Lenovo
Microsoft	MEC	MSN	Staples
Paypal	MSNWeather	9292.nl	Rijksmuseum.nl

Extraction of CSS styles and DOM states. Mazinanian et al. used the dynamic analysis features of Crawljax [3] to dynamically capture different DOM states of a web application. These DOM states are persisted to HTML files.

The HTML files contain inline and internal style sheets, together with links to external style sheets. In order to extract the external style sheets Mazinanian et al. developed an external CSS file extractor plug-in for Crawljax.

Note that the references to the external style sheets in the HTML documents need to be updated, because the extracted external style sheets are in a different location.

Issues with dataset. There are some problems with the dataset of Mazinanian et al., which we fixed by:

- renaming "Apple.ca" to "Apple.com" and "MountainEquip" is "MEC";
- removing a redirect from the Gmail system code which was fired in the absence of a session cookie and recapturing the intended state of the Gmail system again (these two subjects are named "Gmail original" and "Gmail fixed");
- refetching 18b91843bb4bcb07c2ba68a01bbb8a02b9eb4c-50.css and 54c660b14dd08ca6b408f07de1f5080d251a4ef2.css from the "About.com" system, which were empty in the original dataset;
- similarly refetching the files d8e76b82abbae61a-4a89fb4000324c09b6413719.css and f20bbdc283941382159c-1e12d655f54f1bdc68c2.css from the "Alibaba" subject;
- excluding unused files the files: 167d8fb4-7eb42d1f908ba5d4141a34f4333b18c9.css, a0f24af3ff23a278-289a1f50a5d6b6598a76415a.css, b5f14f865786216f67ae8a-41ab1c1774aa955334.css, c7a368297aab3abe7d74e0ae421f-c38dd18a048f.css and e54053a51b1eb8ae62e9bc76ad9351ea-4f1d4c89.css from "ProToolsExpress";
- removing the file style.css from "SyncCreative" since this file is a duplicate of b4ad21b4c1ba9945-1234f5e3da9a501a50dac0fe.css.

We have decided to leave three files empty in order to stay close to the original dataset: 91cab95bff78fd3800625c07-89cd87c0e4180299.css from "Apple.com", 61dc696007ca-3c1aeb54a2d0bab8ea932de60e21.css from "GlobalTVBC" and 0e6fedfab56593cd6d0ea0bd8dee80454585b2af.css from "SyncCreative". The first one returned a File Not Found error and the last two returned a Forbidden error while crawling. However, we do not know if these were also the errors that occurred when the original dataset was collected.

References

[1] D. Mazinanian, "Dataset for FSE'14 submission." [Online]. Available: http://users.encs.concordia.ca/~d_mazina/papers/FSE'14/

[2] D. Mazinanian, N. Tsantalis, and A. Mesbah, "Discovering refactoring opportunities in cascading style sheets," in *Proceedings of the ACM SIGSOFT International Symposium on the Foundations of Software Engineering (FSE)*, 2014.

[3] A. Mesbah, A. van Deursen, and S. Lenselink, "Crawling Ajax-Based Web Applications Through Dynamic Analysis of User Interface State Changes," *ACM Transactions on the Web*, vol. 6, no. 1, pp. 3:1–3:30, 2012.

[4] L. Punt, S. Visscher, and V. Zaytsev, "The A?B*A Pattern: Undoing Style in CSS and Refactoring Opportunities It Presents," in *Proceedings of the 32nd International Conference on Software Maintenance and Evolution (ICSME)*, 2016.

Where to find it

- **Description**
  - DOI [10.1109/ICSME.2016.91](https://doi.org/10.1109/ICSME.2016.91)
  - page 616
  - © IEEE
- **Dataset**
  - http://leonardpunt.github.io/masterproject/dataset-and-results.zip
  - MIT License

Experimental Data for the A?B*A Pattern in CSS: Inputs and Outputs

@leonardpunt
Leonard Punt
University of Amsterdam, The Netherlands
Q42, The Netherlands

@sjoerdvisscher
Sjoerd Visscher
Q42, The Netherlands

@grammarware
Vadim Zaytsev
University of Amsterdam, The Netherlands
Raincode, Belgium

The complete dataset is available under the MIT license from this web location: http://leonardpunt.github.io/masterproject/dataset-and-results.zip. This dataset is used to detect undoing style in CSS code. In total, this dataset contains 41 subjects. Each subject has its own folder, which contains the captured states, a `states.html` file, is used to load all captured states in one document, and a folder called `results`, which contains the detected undoing styles, the refactored style sheets and the detected semantic changes. The file `states.html` was used as an input for our detection tool [4].

Selection of subjects. In order to select representative real-world web applications, we used the empirical data used in the study conducted by Mazinanian et al. [2]. This data set includes 38 popular online web applications and is available online [1]. Besides the 38 subjects from the study of Mazinanian et al., two web applications developed by Q42 have been studied as well. The 41st subject is a new version of the subject 'Gmail' from in the original data set. The complete list of selected systems follows:

Facebook	Pinterest	YouTube	Reddit
Twitter	Tumblr.com	YahooMail	Wordpress.org
Outlook.com	Vimeo.com	Gmail	Igloo
Github	Phormer	Amazon.ca	BeckerElectric
Ebay	Equus	About.com	ProToolsExpress
Alibaba	UniqueVanities	Apple.com	ICSE12
BBC	EmployeeSolutions	CNN	SyncCreative
Craiglist	GlobalTVBC	Imgur	Lenovo
Microsoft	MEC	MSN	Staples
Paypal	MSNWeather	9292.nl	Rijksmuseum.nl

Extraction of CSS styles and DOM states. Mazinanian et al. used the dynamic analysis features of Crawljax [3] to dynamically capture different DOM states of a web application. These DOM states are persisted to HTML files.

The HTML files contain inline and internal style sheets, together with links to external style sheets. In order to extract the external style sheets Mazinanian et al. developed an external CSS file extractor plug-in for Crawljax.

Note that the references to the external style sheets in the HTML documents need to be updated, because the extracted external style sheets are in a different location.

Issues with dataset. There are some problems with the dataset of Mazinanian et al., which we fixed by:

- renaming "Apple.ca" to "Apple.com" and "MountainEquip" is "MEC";
- removing a redirect from the Gmail system code which was fired in the absence of a session cookie and recapturing the intended state of the Gmail system again (these two subjects are named "Gmail original" and "Gmail fixed");
- refetching 18b91843bb4bcb07c2ba68a01bbb8a02b9eb4c-50.css and 54c660b14dd08ca6b408f07de1f5080d251a4ef2.css from the "About.com" system, which were empty in the original dataset;
- similarly refetching the files d8e76b82abbae61a-4a89fb4000324c09b6413719.css and f20bbdc283941382159c-1e12d655f54f1bdc68c2.css from the "Alibaba" subject;
- excluding unused files the files: 167d8fb4-7eb42d1f908ba5d4141a34f4333b18c9.css, a0f24af3ff23a278-289a1f50a5d6b6598a76415a.css, b5f14f865786216f67ae8a-41ab1c1774aa955334.css, c7a368297aab3abe7d74e0ae421f-c38dd18a048f.css and e54053a51b1eb8ae62e9bc76ad9351ea-4f1d4c89.css from "ProToolsExpress";
- removing the file style.css from "SyncCreative" since this file is a duplicate of b4ad21b4c1ba9945-1234f5e3da9a501a50dac0fe.css.

We have decided to leave three files empty in order to stay close to the original dataset: 91cab95bff78fd3800625c07-89cd87c0e4180299.css from "Apple.com", 61dc696007ca-3c1aeb54a2d0bab8ea932de60e21.css from "GlobalTVBC" and 0e6fedfab56593cd6d0ea0bd8dee80454585b2af.css from "SyncCreative". The first one returned a File Not Found error and the last two returned a Forbidden error while crawling. However, we do not know if these were also the errors that occurred when the original dataset was collected.

REFERENCES

[1] D. Mazinanian, "Dataset for FSE'14 submission." [Online]. Available: http://users.encs.concordia.ca/~d_mazina/papers/FSE'14/

[2] D. Mazinanian, N. Tsantalis, and A. Mesbah, "Discovering refactoring opportunities in cascading style sheets," in *Proceedings of the ACM SIGSOFT International Symposium on the Foundations of Software Engineering (FSE)*, 2014.

[3] A. Mesbah, A. van Deursen, and S. Lenselink, "Crawling Ajax-Based Web Applications Through Dynamic Analysis of User Interface State Changes," *ACM Transactions on the Web*, vol. 6, no. 1, pp. 3:1–3:30, 2012.

[4] L. Punt, S. Visscher, and V. Zaytsev, "The A?B*A Pattern: Undoing Style in CSS and Refactoring Opportunities It Presents," in *Proceedings of the 32nd International Conference on Software Maintenance and Evolution (ICSME)*, 2016.

Where to find it

- **Description**
  - DOI [10.1109/ICSME.2016.91](https://doi.org/10.1109/ICSME.2016.91)
  - page 616
  - © IEEE
- **Dataset**
  - http://leonardpunt.github.io/masterproject/dataset-and-results.zip
  - MIT License
- **Ancestor**
  - http://dmazinanian.me/conference-papers/fse/2014/06/16/fse14.html
  - http://dmazinanian.me/publications/FSE'14/FSE'14-crawled.7z

Experimental Data for the A?B*A Pattern in CSS: Inputs and Outputs

@leonardpunt
Leonard Punt
University of Amsterdam, The Netherlands
Q42, The Netherlands

@sjoerdvisscher
Sjoerd Visscher
Q42, The Netherlands

@grammarware
Vadim Zaytsev
University of Amsterdam, The Netherlands
Raincode, Belgium

The complete dataset is available under the MIT license from this web location: http://leonardpunt.github.io/masterproject/dataset-and-results.zip. This dataset is used to detect undoing style in CSS code. In total, this dataset contains 41 subjects. Each subject has its own folder, which contains the captured states, a `states.html` file, is used to load all captured states in one document, and a folder called `results`, which contains the detected undoing styles, the refactored style sheets and the detected semantic changes. The file `states.html` was used as an input for our detection tool [4].

Selection of subjects. In order to select representative real-world web applications, we used the empirical data used in the study conducted by Mazinanian et al. [2]. This data set includes 38 popular online web applications and is available online [1]. Besides the 38 subjects from the study of Mazinanian et al., two web applications developed by Q42 have been studied as well. The 41st subject is a new version of the subject 'Gmail' from in the original data set. The complete list of selected systems follows:

Facebook	Pinterest	YouTube	Reddit
Twitter	Tumblr.com	YahooMail	Wordpress.org
Outlook.com	Vimeo.com	Gmail	Igloo
Github	Phormer	Amazon.ca	BeckerElectric
Ebay	Equus	About.com	ProToolsExpress
Alibaba	UniqueVanities	Apple.com	ICSE12
BBC	EmployeeSolutions	CNN	SyncCreative
Craiglist	GlobalTVBC	Imgur	Lenovo
Microsoft	MEC	MSN	Staples
Paypal	MSNWeather	9292.nl	Rijksmuseum.nl

Extraction of CSS styles and DOM states. Mazinanian et al. used the dynamic analysis features of Crawljax [3] to dynamically capture different DOM states of a web application. These DOM states are persisted to HTML files.

The HTML files contain inline and internal style sheets, together with links to external style sheets. In order to extract the external style sheets Mazinanian et al. developed an external CSS file extractor plug-in for Crawljax.

Note that the references to the external style sheets in the HTML documents need to be updated, because the extracted external style sheets are in a different location.

Issues with dataset. There are some problems with the dataset of Mazinanian et al., which we fixed by:

- renaming "Apple.ca" to "Apple.com" and "MountainEquip" is "MEC";
- removing a redirect from the Gmail system code which was fired in the absence of a session cookie and recapturing the intended state of the Gmail system again (these two subjects are named "Gmail original" and "Gmail fixed");
- refetching 18b91843bb4bcb07c2ba68a01bbb8a02b9eb4c-50.css and 54c660b14dd08ca6b408f07de1f5080d251a4ef2.css from the "About.com" system, which were empty in the original dataset;
- similarly refetching the files d8e76b82abbae61a-4a89fb4000324c09b6413719.css and f20bbdc283941382159c-1e12d655f54f1bdc68c2.css from the "Alibaba" subject;
- excluding unused files the files: 167d8b4-7eb42d1f908ba5d4141a34f4333b18c9.css, a0f24af3ff23a278-289a1f50a5d6b6598a76415a.css, b5f14f865786216f67ae8a-41ab1c1774aa955334.css, c7a368297aab3abe7d74e0ae421f-c38dd18a048f.css and e54053a51b1eb8ae62e9bc76ad9351ea-4f1d4c89.css from "ProToolsExpress";
- removing the file style.css from "SyncCreative" since this file is a duplicate of b4ad21b4c1ba9945-1234f5e3da9a501a50dac0fe.css.

We have decided to leave three files empty in order to stay close to the original dataset: 91cab95bff78fd3800625c07-89cd87c0e4180299.css from "Apple.com", 61dc696007ca-3c1aeb54a2d0bab8ea932de60e21.css from "GlobalTVBC" and 0e6fedfab56593cd6d0ea0bd8dee80454585b2af.css from "SyncCreative". The first one returned a File Not Found error and the last two returned a Forbidden error while crawling. However, we do not know if these were also the errors that occurred when the original dataset was collected.

REFERENCES

[1] D. Mazinanian, "Dataset for FSE'14 submission." [Online]. Available: http://users.encs.concordia.ca/~d_mazina/papers/FSE'14/

[2] D. Mazinanian, N. Tsantalis, and A. Mesbah, "Discovering refactoring opportunities in cascading style sheets," in *Proceedings of the ACM SIGSOFT International Symposium on the Foundations of Software Engineering (FSE)*, 2014.

[3] A. Mesbah, A. van Deursen, and S. Lenselink, "Crawling Ajax-Based Web Applications Through Dynamic Analysis of User Interface State Changes," *ACM Transactions on the Web*, vol. 6, no. 1, pp. 3:1–3:30, 2012.

[4] L. Punt, S. Visscher, and V. Zaytsev, "The A?B*A Pattern: Undoing Style in CSS and Refactoring Opportunities It Presents," in *Proceedings of the 32nd International Conference on Software Maintenance and Evolution (ICSME)*, 2016.

Replications are tough

RAINCODE

- **Mesbah, van Deursen, Roest, “Invariant-Based Automatic Testing of Modern Web Applications”, IEEE TSE 38(1), 2012.**
  - Sven Rohde, “Test-suite for Automated Invariant-Based Testing of Ajax Web-Applications”, UvA SE MSc thesis, March 2015.
  - http://www.scriptiesonline.uba.uva.nl/533688

Replications are tough

RAINCODE

- **Mesbah, van Deursen, Roest, “Invariant-Based Automatic Testing of Modern Web Applications”, IEEE TSE 38(1), 2012.**
  - Sven Rohde, “Test-suite for Automated Invariant-Based Testing of Ajax Web-Applications”, UvA SE MSc thesis, March 2015.
  - http://www.scriptiesonline.uba.uva.nl/533688
- **Mesbah, Mirshokraie, “Automated Analysis of CSS Rules to Support Style Maintenance”, ICSE, 2012.**
  - Kevin Adegeest, “Automated Detection of Unused CSS Style Rules by Crawling Web Applications”, UvA SE MSc thesis, August 2015.
  - http://www.scriptiesonline.uba.uva.nl/588353

Replications are tough

RAINCODE

- **Mesbah, van Deursen, Roest, "Invariant-Based Automatic Testing of Modern Web Applications", IEEE TSE 38(1), 2012.**
  - Sven Rohde, "Test-suite for Automated Invariant-Based Testing of Ajax Web-Applications", UvA SE MSc thesis, March 2015.
  - http://www.scriptiesonline.uba.uva.nl/533688
- **Mesbah, Mirshokraie, "Automated Analysis of CSS Rules to Support Style Maintenance", ICSE, 2012.**
  - Kevin Adegeest, "Automated Detection of Unused CSS Style Rules by Crawling Web Applications", UvA SE MSc thesis, August 2015.
  - http://www.scriptiesonline.uba.uva.nl/588353
- **Mazinanian, Tsantalis, Mesbah, "Discovering Refactoring Opportunities in Cascading Style Sheets", FSE, 2014.**
  - Leonard Punt, "The A-B*-A Pattern of Undoing Style in Cascading Style Sheets", UvA SE MSc thesis, August 2015.
  - http://www.scriptiesonline.uba.uva.nl/588297

Google

Laurence, What the World's Biggest Websites Looked Like at Launch, 2012.
https://www.carocreative.uk/what-the-worlds-biggest-websites-looked-like-at-launch/

Twitter

Laurence, What the World's Biggest Websites Looked Like at Launch, 2012.
https://www.carocreative.uk/what-the-worlds-biggest-websites-looked-like-at-launch/

Speed of deployment

RAINCODE

"At Etsy 150 engineers deploy a single monolithic application more than

60 times a day"

Schauenberg, Development, Deployment & Collaboration at Etsy, QCon London, 2014,
https://qconlondon.com/london-2014/london-2014/presentation/Development,%20Deployment%20&%20Collaboration%20at%20Etsy.html

RAINCODE

21900

small revisions

per year

Details

- Available at
  - http://leonardpunt.github.io/masterproject/dataset-and-results.zip
- Size
  - 23.4 MB packed
  - 171 MB unpacked

Details

RAINCODE

- Available at
  - http://leonardpunt.github.io/masterproject/dataset-and-results.zip
- Size
  - 23.4 MB packed
  - 171 MB unpacked
- 41 directories
  - index.html
  - states.html
  - css/*.css
  - results/*
- 38 sites from Mazinanian-Tsantalis-Mesbah, +2 proprietary, +1

Issues

- **Mostly inherited everything**

Issues

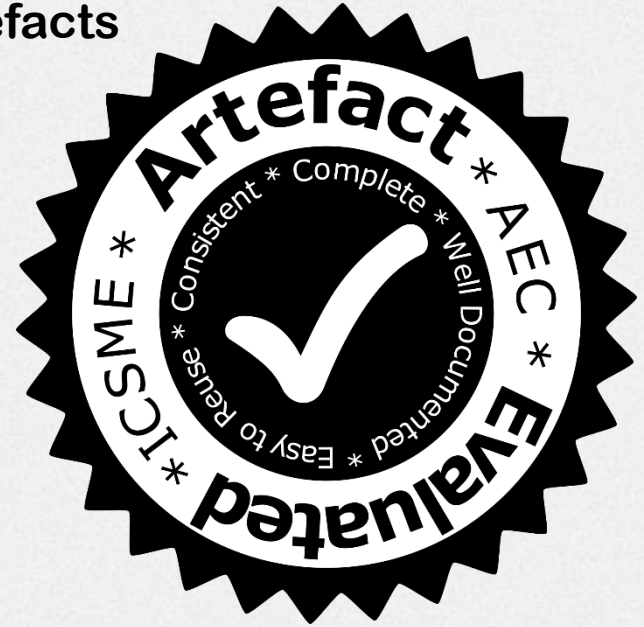
- Mostly inherited everything
- Followed the same *crawling* process for new artefacts

Issues

- Mostly inherited everything
- Followed the same *crawling* process for new artefacts
- Renamed Apple.ca and MountainEquip

Issues

- Mostly inherited everything
- Followed the same *crawling* process for new artefacts
- Renamed Apple.ca and MountainEquip
- Gmail problems (missing cookies)



Issues

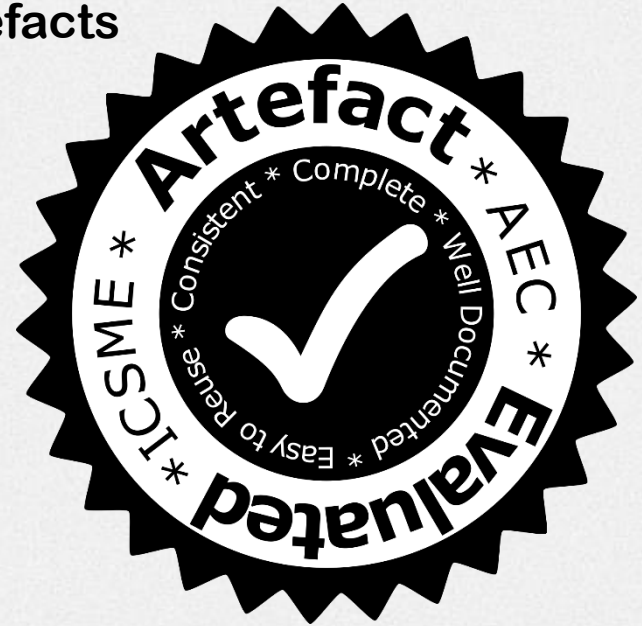
- Mostly inherited everything
- Followed the same *crawling* process for new artefacts
- Renamed Apple.ca and MountainEquip
- Gmail problems (missing cookies)
- File Not Found errors

Issues

- Mostly inherited everything
- Followed the same *crawling* process for new artefacts
- Renamed Apple.ca and MountainEquip
- Gmail problems (missing cookies)
- File Not Found errors
- Forbidden errors

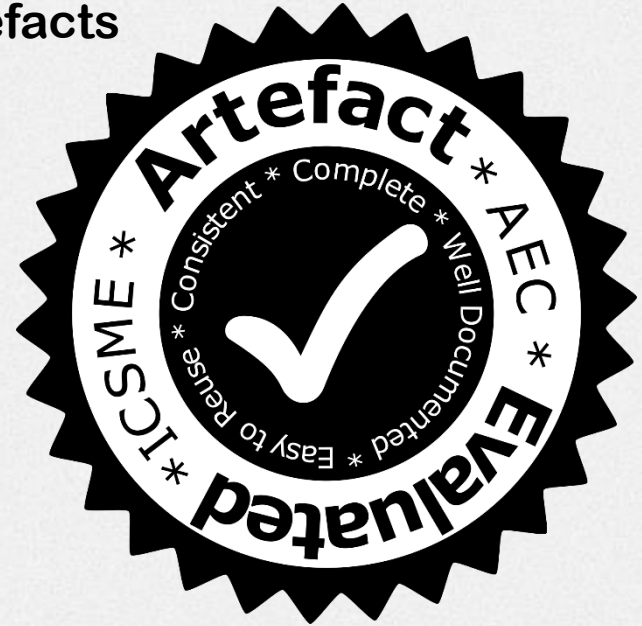
Issues

- Mostly inherited everything
- Followed the same *crawling* process for new artefacts
- Renamed Apple.ca and MountainEquip
- Gmail problems (missing cookies)
- File Not Found errors
- Forbidden errors
- Empty yet neither Forbidden nor Not Found



Issues

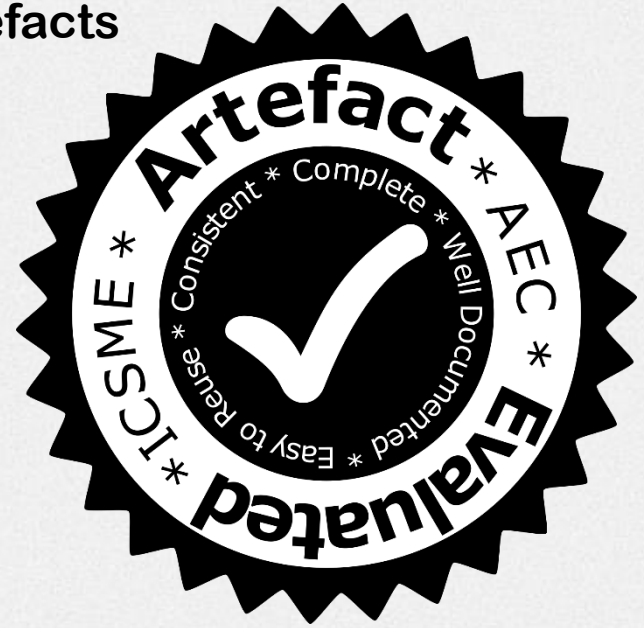
- Mostly inherited everything
- Followed the same *crawling* process for new artefacts
- Renamed Apple.ca and MountainEquip
- Gmail problems (missing cookies)
- File Not Found errors
- Forbidden errors
- Empty yet neither Forbidden nor Not Found
- Unused files



Issues

RAINCODE

- Mostly inherited everything
- Followed the same *crawling* process for new artefacts
- Renamed Apple.ca and MountainEquip
- Gmail problems (missing cookies)
- File Not Found errors
- Forbidden errors
- Empty yet neither Forbidden nor Not Found
- Unused files
- Duplicate files



Alternative use

RAINCODE

- **August 2016: Compiler Construction Summer school**
  - CSS as a running example

Alternative use

RAINCODE

- August 2016: Compiler Construction Summer school
  - CSS as a running example
- https://github.com/software-engineering-amsterdam/ccss/tree/master/2016/grammarware/csss
- Test suite
  - 345 unique CSS files

Alternative use

RAINCODE

- August 2016: Compiler Construction Summer school
  - CSS as a running example
- https://github.com/software-engineering-amsterdam/ccss/tree/master/2016/grammarware/csss
- Test suite
  - 345 unique CSS files
- Day 1: Learn Rascal
- Day 2: Parse as many files as possible
- Day 3: Write metaprograms…

Alternative use

RAINCODE

- August 2016: Compiler Construction Summer school
  - CSS as a running example
- https://github.com/software-engineering-amsterdam/ccss/tree/master/2016/grammarware/csss
- Test suite
  - 345 unique CSS files
- Day 1: Learn Rascal
- Day 2: Parse as many files as possible
- Day 3: Write metaprograms…

- Conclusion: 10-100 files parsed unambiguously
- Student assistants went up to 300.

Conclusions

- **Tools are not everything**
  - Respect your data

Conclusions

RAINCODE

- **Tools are not everything**
  - Respect your data
- **Web apps are way too liquid**
  - Like mining software repositories
  - …without versioning

Conclusions

RAINCODE

- Tools are not everything
  - Respect your data
- Web apps are way too liquid
  - Like mining software repositories
  - …without versioning
- Thanks to Davood, Nikolaos and Ali!

Conclusions

RAINCODE

- Tools are not everything
  - Respect your data
- Web apps are way too liquid
  - Like mining software repositories
  - …without versioning
- Thanks to Davood, Nikolaos and Ali!
- Workflow
  - Fixed
  - Extended
  - Added results
  - Republished

IEEE 32nd International Conference
on Software Maintenance and Evolution (ICSME 2016)

BEST ARTIFACT AWARD

Presented to

Leonard Punt, Sjoerd Visscher and Vadim Zaytsev

Co-Authors of:
"The A?B*A Pattern: Undoing Style in CSS
And Refactoring Opportunities in Presents"

5-7 October 2016

Nicholas A. Kraft
General Co-Chair

T.J. Menzies

Tim Menzies
General Co-Chair