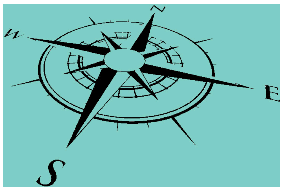
Technological Approaches to Improving Credibility Assessment on the Web: Assessment Strategies

This document is a primer for developing and improving technological methods to help promote trust and accuracy, especially on the web and involving news reporting. While not always comprehensive, it attempts to guide people away from overly simplistic designs and reveal a wide array of potential solutions. It concludes by enumerating technical standards that will be needed to enable many of these methods on the open web. [This StratML rendition documents the assessment strategies as goals to be pursued.]

This is an Internal Review Draft (an "Editors Draft")... This report, like the Community Group which developed it, is focused on web-centric technical solutions which require standardization.

Contents



Credible Web Community Group (CWCG)

Description:

The Credible Web Community Group was formed at W3C, the organization which develops technical standards for the web, to look for technological approaches to this “credibility assessment” problem. It's not that we think technology can solve every problem, especially ones as deeply human and complex as this one, but it seems likely that some technology is making matters worse and that certain designs could probably serve people better. For some of us, creating better approaches to credibility assessment seems like a good way to help.

Stakeholder(s):

Sandro Hawke :
Editor

W3C

Consumers :
(Content) Person who is receiving and experiencing some content. Similar to: Audience Member, Reader, Viewer, Listener, Receiver, or User.

Providers :
(Content) Provider or Source. Person or organization who provided the consumer with some content. There may be a supply chain of providers, creating and assembling content before it reaches the consumer. Alternatively, sometimes a chain of friends act as provider, passing the content on to each other. Often this is invisible to the consumer, who perceives (and makes credibility assessments about) a single apparent provider. Similar to: Producer, Creator, Author, or Publisher. “Source” is often more ambiguous, but can be used as an alternative when “provider” gets awkward.

Promoters :
(Content) Person or organization who intentionally or unintentionally increases the spread of content. In social media platforms, this can be as simple as “liking.” Commenting on content or linking to it, even to refute it, can increase its spread and visibility due to various algorithms.

Facilitators :
(Credibility) Person or organization who is helping the consumer decide what to trust. Similar to: Moderator, Fact-Checker, Forum, or Comments Editor, but also includes members of the crowd in crowd-source designs.

Credibility Stakeholders :
Who Might Care Enough To Do Something

W3C Verifiable Claims Working Group

Schema.org

IEEE-P7011

The Trust Project :
Sally Lehrman

International Fact Checking Network (IFCN) :
at Poynter

Certified Content Coalition (CCC) :
of CableLabs, Scott Yates

Journalism Trust Initiative (JTI) :
initiated by Reporters Without Borders (RSF)

Scientific Researchers :
There are dozens if not hundreds of research groups studying aspects of credibility around the world. Listed here are groups that have expressed interest in helping with technical standards work around credibility.

Credibility Coalition (CredCo) :
led by Meedan and Hacks/Hackers

Haystack Group :
at MIT CSAIL, David Karger

Duke Reporters' Lab

First Draft News :
(at Harvard Kennedy School, Shorenstein Center), Claire Wardle

Center for Complex Networks and Systems Research :
at Indiana University Bloomington

Tow-Knight Center for Entrepreneurial Journalism :
at the Craig Newmark Graduate School of Journalism at the City University of New York (CUNY)

Data & Society

— continued next page

Stakeholders (continued)

Center for Civic Media :
at the MIT Media Lab, Ethan Zuckerman

German Research Center for Artificial Intelligence (DFKI) :
Georg Rehm

UK VIsual Social Media Lab :
Farida Vis

Grant-Makers

News Integrity Initiative :
(at CUNY)

Misinformation Solutions Forum :
(at Rita Allen Foundation)

The Knight Commission on Trust, Media and Democracy :
(at Aspen Institute)

Craig Newmark Philanthropies

Credibility Coalition :
(microgrants for indicator research)

Product Developers :

Hypothes.is :
Web annotations

Meedan News :
verification tools

AboutThem.info :
plans to facilitate applications based on me/us/them statements published on the Web in Strategy Markup Language (StratML) format.

Industries :
Specific vendors/products are named as an example when an employee or representative has joined the group.

News Feed Providers :
News feeds, social media (Facebook)

Social Media

Web Search Providers :
(Google, Bing)

Content Hosting Services :
(YouTube)

Content Aggregators :
Content aggregators & portals

Content Portals

Web Browsers :
(Chrome, Firefox, IE/Edge)

Multi-Sided Markets :
especially with individuals (Airbnb)

Governments

Scientific Research Institutions :
(for communications/outreach) (LBL)

Internet Governance :
(nic.br)

Content Publishers :
(BBC, Wiley)

Content Publisher Associations :
(AP)

Fact Checkers :
(Snopes)

Advertisers

Ad Networks :
(AppNexus)

Credibility Tools :
(Meedan, Hypothes.is, FactsMission)

Vision

The credibility of information on the Web is more easily evaluated

Mission

To provide a primer for developing and improving technological methods to help promote trust and accuracy

Values

Credibility: Credibility (of information, for a given set of consumers). Degree to which information is credible; degree to which information appears non-misleading and useful (for the given audience). People are typically misled when falsehoods have high credibility (appear true), and they resist believing facts which have low credibility (appear untrue). The term credibility can also be used broadly to refer to the problem space around trust, as in "credibility researchers" or "credibility software.

Standardization: With standardization, we can help people (via their computers) work together toward improving credibility assessment and the availability of trustworthy information. As detailed below in Potential New Web Standards, the area we find most suitable for standardization is web data interchange, where websites publish data for other systems to consume. W3C standards already cover the basic mechanics for this, and they are widely implemented, most popularly with vocabularies defined by schema.org (a collaboration among search engine companies). What remains is to standardize vocabularies (also known as schemas or ontologies) for exchanging data which bears directly on credibility assessment, that is, “credibility data.” There may also be a benefit in standardizing new browser functionality. For example, browsers might manage a collection of independent credibility assessment tools which work together to guide and inform the user. This kind of new feature can be pioneered in browser extensions and then later adopted into browsers to increase the user base.

Trust: Trustworthy and Trustworthiness can be synonyms of credible and credibility, respectively, but there seems less consensus on their meaning. For example, a counterfeit that fools everyone is misleading, and if it fools everyone we know it appears non-misleading, so it is “credible”. People seem to disagree about whether such a counterfeit would be considered “trustworthy”.

Accuracy: Degree of truth, especially for claims involving measurements. Can be a useful concept to avoid pedantic distinctions around truth. The claim, “The radius of the earth is 6400 km,” is only somewhat true, as 6371 km is a more accurate figure. Of course, it is not exactly 6371 km, either. Similarly, most “true” statements are not perfectly and completely true, so considering their accuracy may be more helpful.

No Harm: First, Do No Harm: Hazards of Intervention -- Changing the technology of the web to empower users to be more accurately informed, while likely to be difficult, sounds laudable and even overdue. But there are likely to be unintended consequences resulting from these changes.

Privacy: Some credibility solutions may impact privacy. For example, a browser plug-in which checks every page the user visits using its proprietary cloud-based service could potentially misuse or leak information. If designed without proper safety measures, it could easily leak private URLs the user visits, and in some cases could even leak secret page contents, such as the user’s medical communications or their employer’s proprietary data.

Factuality: Fact. A claim that is true; a claim that accurately describes the actual state of the world.

Beliefs: Believe (in a claim, as in: “Alice believes the world is round”). To be in the mental state of accepting (and consequently behaving as if) the claim is true, at least within some limited context. Evidence suggests people can sometimes believe contradictory claims and switch among them depending on the context. Trusting a claim is essentially synonymous, although perhaps a bit stronger, like “firmly believing.” One approach to measuring strength of belief is asking how much someone would bet on the claim being true. People routinely bet their lives on their trust in the safety of products like cars and medicines in return for minor gain. The following definitions are expressed in terms of “some information.” This idea can be applied narrowly, to a small bit of text, or very broadly, to all the information provided over time by some content provider. For example, at small scale, we can consider the credibility of a particular sentence, or, at large scale, the credibility of a particular news organization in a given month. See additional discussion below in Granularity.

Decision Making: Making credibility decisions, the act of deciding for oneself what to believe, is often informal, immediate, and unconscious. Doing this incorrectly often leads to being misled, although it may not be practical to do it correctly at all times. Our hope is to help people do this better, in practice. (See Mike Caulfield's "four moves" for a description of a formal strategy for consciously delaying judgment.)

Analyses: Credibility Analysis, by people or machines, is gathering, organizing, analyzing evidence to help people make credibility decisions about a particular information item. A credibility analysis process might produce some kind of report which might itself by called a credibility assessment and might include a credibility score.

1. Inspection

Look closely at the content (and the page where it appears) for the presence of features which are associated with low or high credibility.

The expectation is that most of these features will be reliably detectable using software soon. Inspection may be the primary strategy consumers use unconsciously while reading. The hope is that assessment accuracy can be significantly improved with more refined software and with additional user studies.

1.1. Tools

Help humans inspect content for credibility signals.

Tools can help humans inspect content for credibility signals. This might take the form of a checklist/questionnaire, combined with an annotation tool, where the user selects the portions of the content which have certain features.

1.2. Annotations

Share annotation data.

That feature annotation data can then be shared, subject to privacy considerations, for multiple purposes.

1.3. Comparison

Compare annotations.

Data from multiple annotators can be compared against each other and against standardized test data to assess how well the measurement process is working, in terms of both precision and accuracy, and to reduce the effects of bad faith annotators.

1.4. Validation

Validate the connection between signals and the truth.

Annotations on test content can be used to validate the connection between signals and the truth. Note the connection might be a positive correlation (signal indicates high credibility), negative correlation (signal indicates low credibility), or there might be a multifactor connection found by machine learning techniques.

1.5. Signaling

Signal credibility levels to other consumers.

This data can be used directly to signal credibility levels to other consumers. It can also be used to train machine models which can then potentially also do the job listed in step 1.

Stakeholder(s):

Consumers

2. Corroboration

Identify claims and check other sources.

Identify salient claims made (or implied, or relied upon) in the content and check other sources which offer an assessment of the claim (e.g., fact-checks), related claims, or evidence helping the consumer accurately assess the claim. This is perhaps the technique most people fall back to when they become suspicious, but it can be prohibitively time-consuming to do regularly or do thoroughly.

2.1. Fact Checking

Allow professional fact-checks to be published in machine-readable form so they can more easily be matched, used in automated assessments, and shown to users when appropriate.

(Already being done using ClaimReview.)

Stakeholder(s):
Fact Checkers

2.2. Claims

Allow claim-extraction processes to publish their output for broad consumption.

This would be a feed of claims found in the media and judged to be of relatively high value for checking.

2.3. User Requests

Allow users to express their desire for specific claims in some content to be checked.

Might include an offer to pay for results.

2.4. Relationships

Allow relationships between claims to be expressed

Allow relationships between claims to be expressed, such as rephrasings to be more precise and context-free versus more terse and context-sensitive; more in agreement versus disagreement with the claim itself; making a broader versus narrower claim, between natural languages, and opposites.

2.5. Structure

Allow structure of a fact-check to be exposed as machine-readable, showing argumentation and secure links to the evidence.

2.6. User Expressions

Allow end-users to express what they feel or know relevant to a claim’s accuracy in a way that might be aggregated into an accurate credibility assessment.

It's important this not be simply popularity polling; important ideas, when they start, are usually believed by exactly one person.

2.7. Sources

Make it easy to view reviews of related claims from many sources at once, including some data about the sources, such as assessed bias.

3. Reputation

Assess the credibility of a content provider.

Stakeholder(s)
Content Providers

Assess the credibility of a content provider by gathering statements about them from other providers, typically ones with known credibility. This can be done recursively, forming a reputation network from a few known and trusted "root" providers to help assess many unknown and suspect ones. This can also help establish the credibility of corroboration sources, and can be done using a combination of institutional data sources (eg certification authorities) and informal/personal data sources, eg the user's social group of friends, contacts, and influencers.

3.1. Identity

Make the identity of the content provider, including any provenance chain of providers behind them, visible to the consumer

Make the identity of the content provider, including any provenance chain of providers behind them, visible to the consumer, especially in ways which align with user's natural skills at recognizing faces and logos. In particularly, systems could try to avoid presenting confusingly similar brands without clear warning, and indicating how familiar their data indicates the brand is to the user.

Stakeholder(s):

Content Providers

Content Consumers

3.2. Software

Use software to ascertain the reputation of each provider

Software working on your behalf can try to ascertain the reputation of each provider you encounter (eg each website you visit) based on what has been said about them by entities you select as trustworthy, and by intermediate parties in a social graph.

3.3. User Assessments

Enable users to provide their assessment of the reputation of sites

Users could be prompted, or at their own initiative enter their assessment of the reputation of particular sites they visit often, for their contacts unfamiliar with the site to use. This information could also come from institutions willing to make statements about their peers.

3.4. Granularity

Enable credibility assessments at different levels of granularity

Reputation might be a complex (see Granularity above) or a simple one-bit-flag indicating whether the speaker considers the subject to always operate in good faith.

3.5. Seals & Signals

Make trust seals and other positive reputation signals from institutions secure.

Some trust seals (like the one from BBB) are more than just an embedding image that anyone can use, but many are not. Even the more secure ones can be abused to mislead users who do not check them. In contrast, an out-of-band secure solution, for example done by the browser, could be much more secure.

3.6. Rebuttals

Tie claims to contrary claims.

Negative reputation, such a demands for retraction, could be shared and displayed when you visit a site, if it comes from sources you trust. With the right software, as a lie travels halfway around the world, it might be firmly tied to its correction, or at least a strong warning about its low credibility.

4. Transparency

Consider what the provider says about themselves and their content.

Stakeholder(s)

Content Providers :

Here, the emphasis shifts toward the provider making an effort to be more visibly trustworthy. Transparency requires the provider reveal information about themselves and their content, which can require considerable effort and often brings significant risk, in the hope of being more credible.

For instance, content may be labeled as opinion or satire, and providers may label themselves as partisan or as having processes and controls the consumer views as untrustworthy. By itself, this data is easy to fake, but coupled with corroboration, reputation, and off-line social and legal incentives, it may be quite effective. Transparency includes providing meta-information which reduces the risk consumers will be misled, such as statements specifying the intended audience or disclaiming certain reliability practices or liability. To some degree, transparency allows providers to set the standards by which they are to be judged, simplifying the credibility assessment task.

4.1. Labeling

Label types of information

Label types of information, particularly information which is not intended to be accurate but could be mistaken for fact (such as parody, fiction, and opinion which is not phrased as an opinion). This is partially done in schema.org's article types. These types could be conveyed to users by the platforms.

4.2. Disclosures

Disclose ownership, funding, and control information

Providers can disclose ownership, funding, and control information about themselves, which can be connected with reputation network information.

Stakeholder(s):

Content Providers

4.3. Highlighting

Highlight which sites have made disclosures

Software can simply highlight which sites have made disclosures

4.4. Aggregation & Analyses

Aggregate and analyze machine-readable disclosures.

Software can aggregate and analyze machine-readable disclosures. This could benefit considerably from standardized optional disclosure clauses, in the style of Creative Commons.

4.5. Relevance

Highlight aspects of the transparency disclosures which might be especially relevant.

Software can highlight aspects of the transparency disclosures which might be especially relevant to you, based on your values or stated trust concerns Software can particularly highlight supportive and refutive claims from known 3rd parties about self-descriptive claims. (Combines with reputation)