

Chapter 12

The WIQA Browser

The WIQA browser is an example application that uses the WIQA framework. The browser demonstrates how information quality filtering capabilities can be integrated into a standard web browser. The browser enables users to extract structured information from web pages. Extracted information from different web pages is stored in a local repository and can be browsed, sorted, and searched together. The content of the local repository can be filtered using quality-based information filtering policies. In order to help users to understand the filtering decisions, the browser can display explanations why a piece of information satisfies a selected policy.

The WIQA browser is based on the Piggy Bank extension for the Firefox web browser developed by the SIMILE project at the Massachusetts Institute of Technology [HMK05]. The WIQA browser uses Piggy Bank functionality to extract structured information from web pages and to display and navigate extracted information. The WIQA browser employs NG4J to store information together with provenance meta-information as a set of named graphs. The browser uses the WIQA - Filtering and Explanation Engine to filter stored information and to generate explanations about filtering decisions.

The WIQA browser is available under the terms of the Berkeley Software Distribution license [Reg99]. The browser is distributed as an XPI installation file which integrates the browser into Firefox. The installation file can be downloaded from the WIQA browser website ¹.

The following sections describe the capabilities of the WIQA browser to collect, filter, and display information. The financial information integration scenario outlined in Chapter 7 is used as a running example to explain how an investor can use the browser to collect financial information from different websites and explore collected information using different filtering policies.

¹<http://www.wiwiss.fu-berlin.de/suhl/bizer/wiqa/browser/> (retrieved 09/25/2006)

12.1 Collecting Information

While a user browses the Web, the WIQA browser runs in the background and analyzes the visited web pages. Whenever the browser can extract structured information items from a page, it shows a *data coin icon* in the status bar of the browser, indicating that the user can switch to an *information item view* of the web page. The *information item view* lists all information items that have been extracted from the page. Next to each item is a *save button*. Pressing the button stores the item in the local repository.

Let us assume that our investor is interested in several companies and plans to buy stocks of one of these companies. For deciding on the right company, he would visit several financial information portals and search for news, analyst reports, and discussion forum postings about the companies. Whenever he finds an interesting information item on one of the websites, he will let the WIQA browser extract the information item and save it to his local repository for future reference.

Figure 12.1 illustrates the investor's information collection process. The screenshot on the left shows a Yahoo Finance! page² containing news about Google Inc. The *data coin icon* in the status bar of the browser indicates that the news can be extracted as structured information items. Clicking on the icon switches to the *information item view* of the page shown on the right in Figure 12.1. Each table on the left-hand side of the information item view represents a news item which can be saved into the local repository.

The WIQA browser can extract information from web pages that offer their content, beside of HTML, in an alternative structured format. The browser supports Really Simple Syndication (RSS) [RSS05], RDF/XML [Bec04b], and RDF/N3 [BL98] as alternative content formats.

On websites which do not provide an alternative structured representation, the WIQA browser can invoke screen-scrapers to extract structured information from within a web page's content. The WIQA browser uses the Piggy Bank screen-scraping framework [SIM06a] to transform webpages into RDF. Each screen scraper defines a URL pattern and is invoked whenever a page matching this URL pattern is visited. Screen-scrapers are implemented as XSLT templates [Cla99] or in Javascript. Alternatively, screen-scrapers can be visually defined using Solvent [SIM06b], a tool which enables users to highlight parts of an XHTML page and automatically generates a Javascript screen-scrapers to extract these parts.

The WIQA browser uses the Named Graphs data model to represent information together with provenance meta-information. Whenever the user

²<http://finance.yahoo.com/q?s=Goog> (retrieved 09/25/2006)

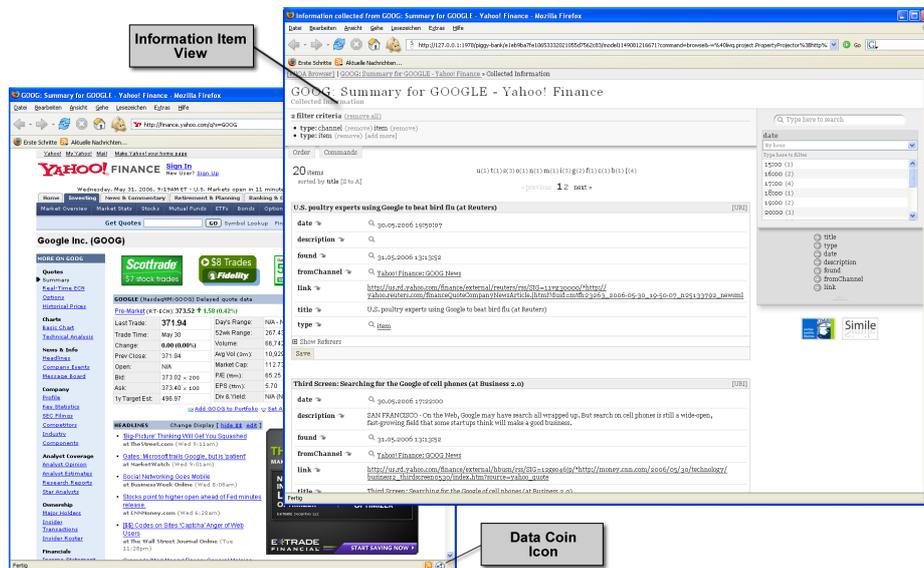


Figure 12.1: Extracting news about Google Inc. from the Yahoo Finance! website.

saves information from a webpage into the local repository, the browser creates a new named graph for this visit of the page and stores the current timestamp, the URL of the page, and the authority (website URL) together with the actual information. The new graph is named with an UUID [LMS05]. Provenance meta-information is represented using the Semantic Web Publishing and the Dublin Core [NPJN06] vocabularies.

Figure 12.2 shows the browser's internal representation of two information items, a person and a document, that have been saved from the FOAF profile [BM04] <http://www.wiwiss.fu-berlin.de/suhl/bizer/foaf.rdf>. Note that the WIQA browser uses the URL of the website from which information is saved as object of the `swp:authority` triple. Screen-scrapers for web pages that contain explicit authorship information can overwrite this URL with a more specific identifier.

12.2 Importing and Exporting Information

In addition to collecting single information items from the Web, the user can also import sets of named graphs into the local repository, and export the content of the local repository to a file. The browser's import and export features are found in the menu *Extras/WIQA Browser* shown in Figure 12.3. The WIQA browser supports the TriX [CS04a], TriG [Biz05],

```

1. @prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2. @prefix dc: <http://purl.org/dc/elements/1.1/> .
3. @prefix swp: <http://www.w3.org/2004/03/trix/swp-2/> .
4. @prefix : <http://www.bizer.de/> .
5.
6. <urn:uuid:8c845860-dce7-11d9-b9c0-00112ff60c7f> {
7.
8.     :i rdf:type foaf:Person ;
9.       foaf:name "Christian Bizer" ;
10.      foaf:mbox <mailto:chris@bizer.de> .
11.
12.     :p747-bizer.pdf rdf:type foaf:Document ;
13.      dc:title "Using Context- and Content-Based ..." .
14.
15.     <urn:uuid:8c845860-dce7-11d9-b9c0-00112ff60c7f>
16.       swp:assertedBy
17.       <urn:uuid:8c845860-dce7-11d9-b9c0-00112ff60c7f> ;
18.       swp:authority <http://www.bizer.de> ;
19.       dc:date "2006-05-14T17:18:10+02:00" ;
20.       swp:savedFrom
21.       <http://www.wiwiss.fu-berlin.de/suhl/bizer/foaf.rdf> .
22. }

```

Figure 12.2: Named Graph generated by the WIQA browser. The graph contains two information items together with provenance meta-information.

RDF/XML [Bec04b], and RDF/N3 [BL98] syntaxes.

Our investor could use the import function to load graph sets that he has received from an information syndicator, like the example graph set from Section 7.2, into the repository. Some filtering policies require specific background information, like the affiliation of information providers, benchmark scores of analysts, or ratings of information providers. The import function can be used to load such background information into the repository. The following examples will assume that the investor has loaded an extended version of the example graph set³ from Section 7.2 into the repository.

12.3 Browsing Information

Figure 12.4 shows the browser's user interface for exploring information in the local repository. Information items from the local repository are displayed on the left-hand side. Each item is rendered as a table, containing

³<http://www.wiwiss.fu-berlin.de/suhl/bizer/wiqa/finUseCase/finData.trig>
(retrieved 09/25/2006)

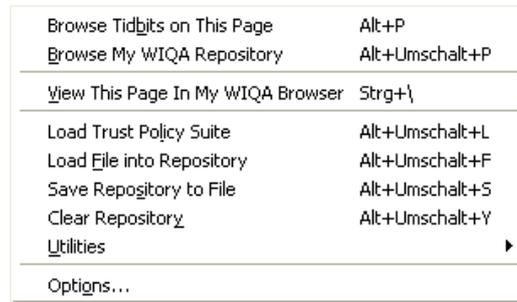


Figure 12.3: The WIQA browser menu.

the item’s properties and their values. The user can refine the collection of items down to a desired subset, by defining filters. Filters are specified using the navigation panel on the right-hand side. There are two types of filters:

Property Filters restrict displayed items to the subset of all items that have a specific property value. For instance, a property filter could restrict items to having the `rdf:type fin:Analyst`, or the `fin:country iso:DE`. Property filters are defined by selecting a property from the *property selection panel*, shown in Figure 12.4, and by selecting the desired property value from the *value selection panel* afterwards.

Text Filters restrict displayed items to the subset of all items that have a property value containing a certain string. A text filter is defined by entering search terms into the *search panel*.

The filter criteria that are currently used are shown above the resulting sub-collection of items. In Figure 12.4, the collection of all items has been zoomed in to the sub-collection of all organizations by applying the property filter “is an Organization”. Clicking *remove* next to a filter criterion would undo the corresponding browsing action and zoom out to the collection of all items.

12.4 Applying Policies and Retrieving Explanations

Back in 1997, Tim Berners-Lee, the inventor of the World Wide Web, envisioned that Web browsers could explain the quality and trustworthiness of displayed information. He proposed that the user interface of each browser should contain an “Oh, yeah?”-button [BL97]. Whenever a surfer does not

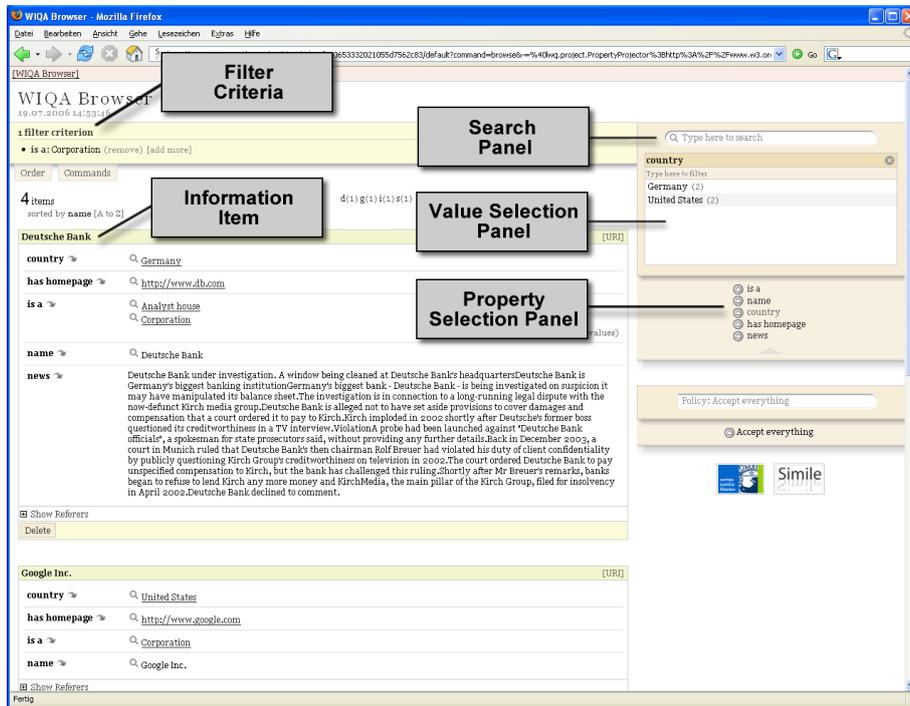


Figure 12.4: Browsing information in the local repository.

feel confident about displayed information, he should press this button and the browser would display an explanation why information should be considered trustworthy.

The WIQA browser realizes the “Oh, yeah?”-button. The user can load a WIQA policy suite into the browser. A policy suite is chosen in the menu *Extras/WIQA-Browser/Load Policy Suite*. The policy suite that is currently loaded is displayed in the *policy selection panel* shown in Figure 12.5. After selecting a policy from the panel, the content of the local repository is filtered using the policy and the left-hand view is updated to show only information matching the policy.

When a policy is applied, an “Oh, yeah?”-button labeled with a question mark appears next to each piece of information. Pressing those buttons opens a new window with an explanation why a piece of information satisfies the selected policy. Figure 12.6 shows an example explanation. The explanation establishes why an analyst report fulfills the policy “Use only information that has been asserted by German analysts”.

The names of resources, such as persons and organizations, are displayed as links in the explanations. Clicking on a link opens up a new browser window containing information about the resource. The user can browse along

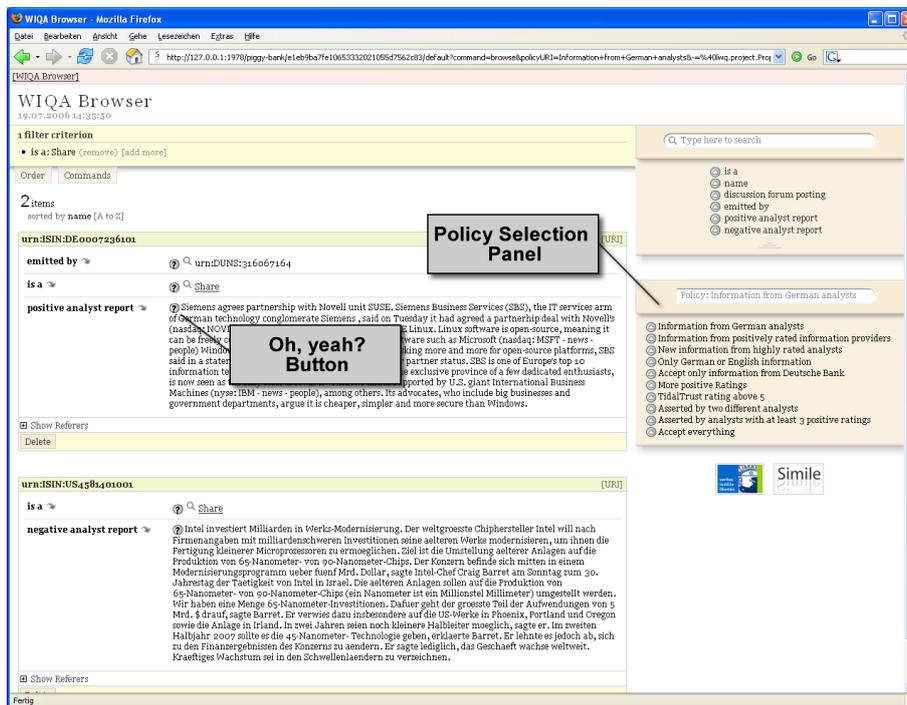


Figure 12.5: When the user selects a policy from the policy selection panel on the right-hand side, the left-hand view updates to show only matching information. The “Oh, yeah?”-buttons open new windows with explanations why a piece of information satisfies the selected policy.

these links to further explore background information. Figure 12.7 shows the browser displaying an explanation together with background information about the analyst Peter Smith.

While browsing, the investor might change filtering policies and select the policies that he feels are appropriate for different tasks. For instance, instead of requiring analyst reports to originate from German analysts, the investor could decide to accept only reports that have been published after a certain date by analysts with a high benchmark score. Figure 12.8 shows an explanation why a report matches this policy.

While browsing discussion forum postings, where the investor does not know information providers directly, he could select a rating-based policy which relies on the Tidal Trust metric. Figure 12.9 shows an explanation why a discussion forum posting satisfies the policy “Accept only information from information providers who have a Tidal Trust score above 5”. The explanation contains the calculation result and details the calculation process.

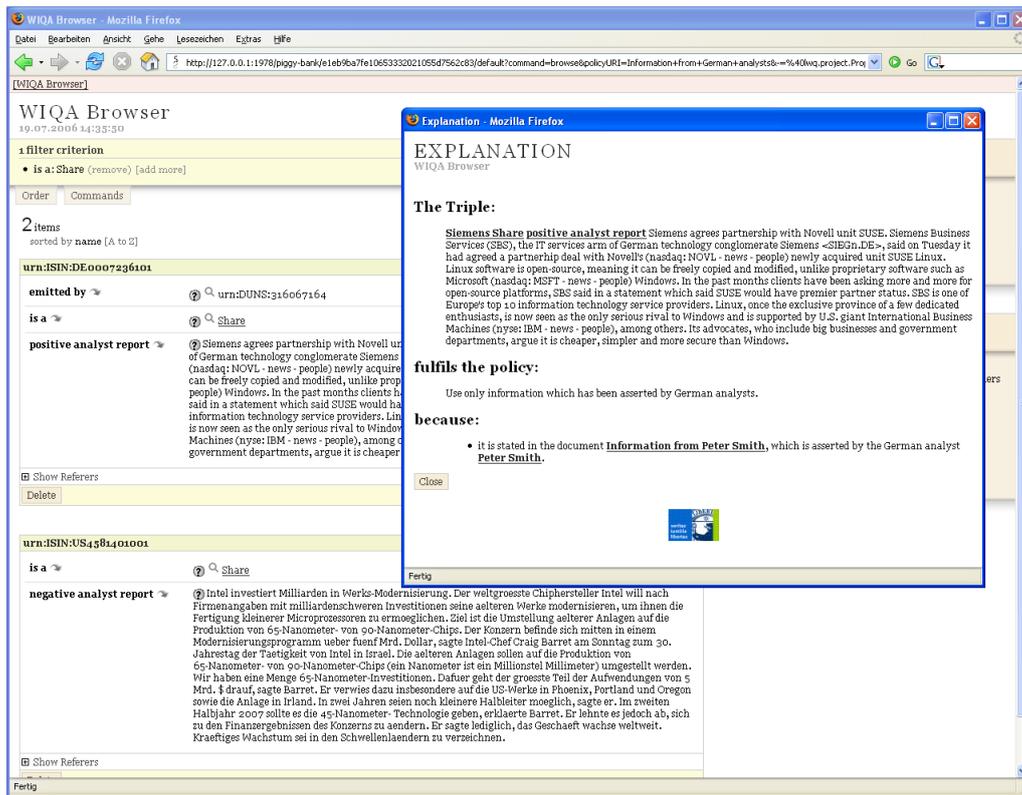


Figure 12.6: The explanation window. This explanation establishes why an analyst report matches the policy “Use only information that has been asserted by German analysts”.

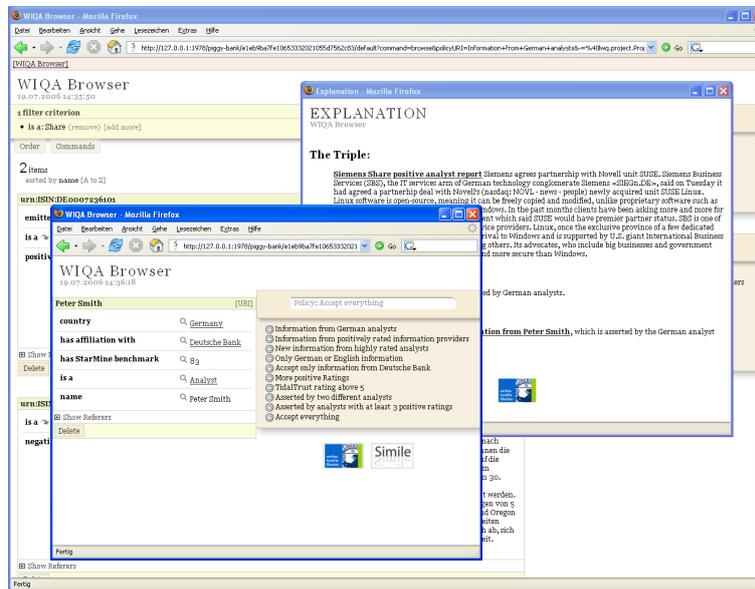


Figure 12.7: Exploring background information: Clicking on the links in the explanation opens a new browser window with background information about a resource.

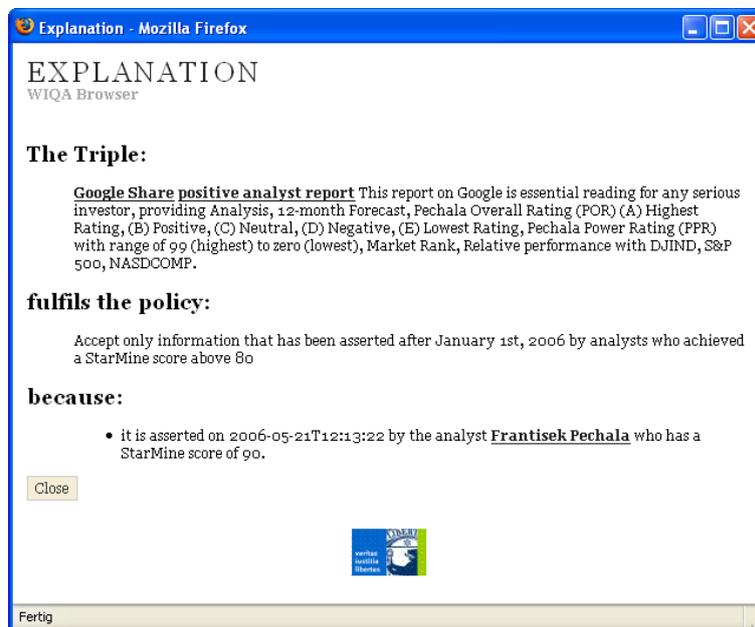


Figure 12.8: This explanation establishes why an analyst report fulfills the policy “Accept only information that has been asserted after January 1st, 2006 by analysts who achieved a StarMine score above 80”.

Explanation - Mozilla Firefox

EXPLANATION

WIQA Browser

The Triple:

Intel Share discussion forum posting As we have already seen in in the past, investing into this company is no good idea.

fulfils the policy:

Only accept information from information providers with a Tidal Trust rating above 5.

because:

- it was asserted by Mark Scott. The WIQA extension function 'Tidal Trust' inferred a rating of 6.7 from Chris Bizer for Mark Scott. (Detail number 1)

Details:

Detail number 1

- The inferred rating arises from the following calculation:
 - The shortest path between Chris Bizer and Mark Scott has length 3. There are 4 different paths with that length. (Detail number 2)
 - The maximum strenght of the paths is 6.0. Therefore, ratings below 6.0 are ignored.
 - The calculation yielded a result of 6.7. (Detail number 3)

Detail number 2

- Paths from the source to the sink:
 - Chris Bizer -6.0-> Anne Richards -9.0-> John Gevner -9.0-> Mark Scott (Strength of the path: 6.0)
 - Chris Bizer -9.0-> Siddhartha Kataki -7.0-> Mary Louis -6.0-> Mark Scott (Strength of the path: 6.0)
 - Chris Bizer -6.0-> Anne Richards -6.0-> Mary Louis -6.0-> Mark Scott (Strength of the path: 6.0)
 - Chris Bizer -6.0-> Anne Richards -5.0-> Ivan Versivic -8.0-> Mark Scott (Strength of the path: 5.0)

Detail number 3

- Chris Bizer -6.7-> Mark Scott was calculated from these ratings:
 - Siddhartha Kataki -6.0-> Mark Scott, weighted with Chris Bizer's rating of 9.0 for Siddhartha Kataki
 - Anne Richards -7.8-> Mark Scott, weighted with Chris Bizer's rating of 6.0 for Anne Richards
 - Anne Richards -7.8-> Mark Scott was calculated from these ratings:
 - Mary Louis -6.0-> Mark Scott, weighted with Anne Richards's rating of 6.0 for Mary Louis
 - John Gevner -9.0-> Mark Scott, weighted with Anne Richards's rating of 9.0 for John Gevner
 - Ivan Versivic -8.0-> Mark Scott, weighted with Anne Richards's rating of 5.0 for Ivan Versivic
 - Siddhartha Kataki -6.0-> Mark Scott was calculated from these ratings:
 - Mary Louis -6.0-> Mark Scott, weighted with Siddhartha Kataki's rating of 7.0 for Mary Louis
 - Ivan Versivic -8.0-> Mark Scott is a direct rating.
 - John Gevner -9.0-> Mark Scott is a direct rating.
 - Mary Louis -6.0-> Mark Scott is a direct rating.

Close

Fertig

Figure 12.9: This explanation establishes why a discussion forum posting fulfills the policy “Accept only information from information providers who have a Tidal Trust score above 5”.