



The Wikirate Project

► D7.7.4 Interim Technical Evaluation 1

D7.7.4

Richard Mills ► Cambridge University ► 20/10/2014

Dissemination level	Public
Contractual date of delivery	Month 12 September 2014
Actual date of delivery	Month 13 October 2014
Work package	WP7 Evaluation
Deliverable number	D7.7.4 Interim Technical Evaluation I
Type	Report
Approval status	Approved
Version	1.1
Number of pages	19
File name	D7.7.4_Interim Technical Evaluation I_20142010_V1.1_Cambridge University

Abstract

This interim report discusses early features of Wikirate in line with T7.2 Technical evaluation issues, and provides recommendations for future development.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.





History

Version	Date	Reason	Revised by
0.5	09/10/2014	Initial draft	Ethan McCutchen
1.0	16/10/2014	Full draft with contributions from Cambridge, CERTH and Wikirate E.v.	Richard Mills
1.1	20/10/2014	Final version	Richard Mills

Author list

Organization	Name	Contact information
Cambridge University	Richard Mills	rm747@cam.ac.uk
GrassCommons	Ethan McCutchen	ethan@grasscommons.org
CERTH	Ioannis Tsampoulatidis	itsam@iti.gr
Wikirate E.v.	Vishal Kapadia	vishal@wikirate.org



Executive Summary

This report presents a high-level overview of the Wikirate project's technical components. It describes the current state of the platform, the direction of future development, and recommendations for improvements.

Particular attention is paid to the question of how Wikirate advances the state of the art. One key point here is the deep atomism of the Wagn platform, which is being augmented and refined as part of the project. From these atoms, Wikirate is building a site which adds valuable structure and process for the crowd-sourced production task.

The Wikirate project will significantly advance the state of the art with respect to peer production in the domain of company information - by breaking this complex task down into granular chunks that are more amenable to a crowd-sourcing approach. In doing so, the Wagn platform's utility is also being improved. These improvements include features which will be of benefit beyond the Wikirate project, like multi-lingual support through user-translation of content,

The Wikirate site will serve as a demonstration that Wagn can be used to build a peer production system which is tailored to a very specific task. The exceptional transparency of both Wagn and Wikirate will facilitate other entities using Wagn to build their own community site which is similarly tailored to their specific use case.



Table of Contents

HISTORY.....	1
AUTHOR LIST.....	1
EXECUTIVE SUMMARY.....	2
TABLE OF CONTENTS.....	3
1 INTRODUCTION.....	4
2 TECHNICAL DEVELOPMENT	4
2.1 SERVER ARCHITECTURE	4
2.2 SCALABILITY	5
2.3 DEVELOPMENT PROCESS.....	7
2.4 TESTING	8
3 INTEGRATION.....	8
3.1 WIKIRATE-CERTH INTEGRATION.....	8
3.2 STRUCTURED SOURCES	9
3.3 SOCIAL MEDIA.....	9
4 ADVANCEMENT OF THE STATE OF THE ART	10
4.1 PRESENT	10
DEEP ATOMISM.....	10
ATOMIC SETS.....	11
COMPREHENSIVE REVISION TRACKING	11
NEW ARCHITECTURAL PATTERN	12
SOURCES AND CLAIMS	12
MULTIMEDIA DUPLICATE DETECTION	13
TOPIC DETECTION.....	13
4.2 FUTURE.....	14
MULTILINGUAL WIKI	14
RATINGS MARKETPLACE	15
DATA QUALITY MECHANISMS	16
EMERGING TOPIC DETECTION	17
DUPLICATE AND TOPIC DETECTION	18
5 RECOMMENDATIONS.....	18



1 Introduction

This report will not undertake the task of detailing year 1 development work; that information is presented in other deliverables (chiefly 6.6.1 and 5.5.1) and annual technical reports. This report will instead present a high-level technical overview of the project from a few key perspectives.

The project's technical infrastructure is described, along with approaches to scalability, the development process and integration of work from different consortium partners. Throughout these sections the current situation is described along with potential improvements which have been identified.

The report also devotes considerable attention to the Wikirate project's broader impact. The ways in which the project advances the state of the art are described in terms of presently deployed features and those which are planned for deployment in the near future.

2 Technical Development

2.1 Server Architecture

The current server architecture for Wikirate.org is quite simple and is comprised of two identical Hetzner EX40 dedicated root servers, one for production and one for development. The servers feature:

- a high-performance Intel® Core i7-4770 Quad-Core processor
- 32 GB DDR3 RAM
- 2 TB SATA 6 Gb/s)
- Network: 620 Gbit Bandwidth

CERTH server (running Solr, MongoDB, MySQL, Apache HTTP, Apache Tomcat, Java, PHP) is comprised of:

- 8x Intel® Core™i7-3770 CPU @ 3.50Ghz
- 32 GB DDR3 RAM
- RAID 2 TB HD
- Network: Node of GR-Net backbone (~230Mbps/download, ~455Mbps/upload)



This is quite appropriate for the site at this stage, and given current traffic projections and server load, all indications are that this architecture will continue to suffice until after Milestone III in month 18. However, a multi-server architecture will be necessary in order for the site to scale to support the traffic envisioned for the full launch at Milestone IV.

The projected future multi-server architecture involves separation of the roles into data servers, app servers, and http servers. The Cambridge team is experienced in deployments of this sort, and the Ruby-on-Rails framework from which Wagn is derived is commonly deployed with this pattern, so we are confident that this plan will serve Wikirate well.

Wikirate has had very high uptime (over 99.9%) this year.

2.2 Scalability

While Wagn/Wikirate is currently quite performant, there will need to be more attention to several areas in order to ensure scalability.

Strengths of the current architecture include its robust “card caching” system, which in Wikirate’s case is deployed on top of Memcached. Because of Wagn’s core principle that “everything is a card”, this caching system is central to its performance and allows for a single piece of caching infrastructure to apply to changes throughout the system. On top of this card caching system, there is a rule caching system that makes card-based rules extremely efficient. A single page rendering can involve hundreds of rules, so the efficiency of this system is critical to performance.

There is also additional evidence that Wagn is capable of performing in substantial traffic situations, thanks to the popularity of a Wagn-based site about, of all things, teen wolves. There is currently a popular television show in the US called “Teen Wolf”, and when this show is aired, the site receives tens of thousands of visitors in the span of a few hours. While there is no formal relationship between the maintainers of this site and any member of the Wikirate consortium, the site maintainers have invited Grass Commons to review their site during these peak traffic events. Their site is indeed stressed during the television shows, but Wagn does not bog down and continues rendering pages throughout. Given that the site uses a single shared server of lesser capacity than Wikirate’s dedicated server, this is certainly a positive sign.

However, several improvements both to the Wagn platform and to the Wikirate modifications will be needed to ensure strong performance with the coming features and projected traffic increases.



In the case of Wagn, the greatest needs are:

- Caching of *processed content*, to reduce repeated processing of intra-content references like inclusions and links
- Caching of multi-card content chunks from common perspectives. While page rendering is quite fast on Wagn, there is still far too much real-time rendering for real time uses, particularly on complex pages. By “common perspectives”, we are referring to the fact that the same page may appear different to different users based on permissions, previous activity, role, or (soon) language preferences, but these special considerations should not necessitate re-rendering for users needing default or common page views.
- Allowing public files and images to bypass the permissions checking system. This check is generally considered a feature on Wagn, because unlike most similar systems Wagn has genuine permission checking on all files. The checks are minimal and fast and upon approval the processing responsibilities are restored to the webserver from the application. However, for scaling purposes publicly-viewable files and images should bypass the application altogether.

We would also encourage strong consideration of moving from the current relational database model to a non-relational database like MongoDB. This model is a very comfortable fit for Wagn’s card architecture and WQL (Wagn Query Language), which effectively already makes Wagn itself act like a noSQL database. This measure would go a long way towards buffering Wikirate against the challenges of dealing with massive amounts of data as anticipated.

In addition to these platform-level considerations, Wikirate itself will need significant optimization attention around several coming features, including:

- Dynamic calculation and recalculation of formulaic company measurements as part of the coming ratings system.
- Caching of amalgamated action counts such as the “contributions” measure for Companies, which would include all actions on related claims, sources, articles, and measurements.
- Dynamic data visualizations

In summary, Wikirate is performing well now, and appropriate attention is being given to paving the way for high performance in a future expected to involve high traffic and great quantities of data. However, there will need to be further attention devoted to these issues in the future in order to support the scale of traffic and data anticipated.



2.3 Development Process

The development team is using an agile and transparent process for generating and refining its technology.

In support of deep coordination between the Wagn platform and Wikirate application developers, the teams meet daily for a “stand up” call to give brief reports of current activity, next steps, and any required assistance or coordination.

Our process deeply embraces the idea of code review for training, coordination, refining solutions, and forming stylistic patterns. The primary code review mechanism is GitHub’s “pull request” system.

The code for the Wikirate Project is currently divided among the following main GitHub repositories:

- The Wagn core: <https://github.com/wagn/wagn>
- Wikirate modifications: <https://github.com/wagn/wikirate>
- Wikirate integrations: <https://github.com/wikirate/data-structure>
- Wikirate-CERTH integrations: <https://github.com/MKLab-ITI/wikirate>

GitHub provides great visibility into the activity of each repository and supports a narrative of high development activity that has accelerated through the first year of the grant.

Deployment of approved code is automated using Capistrano, a popular ruby deployment configuration framework.

One clear area for potential improvement in Wikirate’s deployment process is the coordination of changes in card content. Because the use of cards for site configuration blurs the line between data and code, it creates some additional challenges in code deployment where code and data need close coordination. This has been improved through enhancements to our import/export system, which has been integrated into the Capistrano deployment. But this system is still very young and lacks many features, like date tracking and handling deletions. Improvements here would help ensure smooth deployment with minimal oversight.



2.4 Testing

Wagn is currently tested with two main frameworks: Rspec and Cucumber. The Wikirate application is tested with Rspec. Test coverage for both is measured with Simplecov.

Great progress has been made over the course of the year in separating the platform from the application in a way that makes independent testing possible for each. The key need was to release Wagn as a library (a ruby gem) rather than as a Rails application as it was formerly deployed.

Testing coverage for Wagn is now approaching 90%, with coverage for Wikirate at around 70%.

While these are passable statistics, we believe they should be further improved by:

- Increasing test coverage to 99% for both Wagn and Wikirate
- Introducing cucumber tests for the Wikirate application
- Deepening the commitment to test driven development and ensuring that tests are written *prior* to development of new features
- Formalizing automated tests of the Wikirate – CERTH integrations
- Running continuous integration tests (efforts to set up such testing commenced recently)
- Measure times for crawling content from social streams
- Measure times for indexing
- Measure times for duplicate detection (visual and textual)
- Measure times for source recommendations

3 Integration

3.1 Wikirate-CERTH Integration

Wikirate.org has recently introduced its first formal integration with CERTH's data processing in the form of suggested sources. On any given Company+Topic page (internally called an "Analysis" page), there appear multiple lists of "Sources to Consider". One such list is "From the Web", and this is comprised of suggested source pages not yet added by Wikirate community members. These sources may be reviewed by community members and added as sources if deemed relevant – or marked as irrelevant.



These sources are generated by providing CERTH with “streams” of potential sources (such as RSS or twitter feeds), which they then process to determine relevance to Companies and Topics tracked on Wikirate.org. When a user deems a source relevant or irrelevant, this information is fed back to CERTH in order to help improve their suggestions.

Currently, the streams are suggested to CERTH directly by Wikirate team members. In the near future these streams will be proposed and tracked directly on Wikirate.org and delivered to CERTH by webservice.

We will also soon be introducing many other CERTH services, including:

- Suggested additional topic/company tags for existing sources
- Duplicate detection of claims
- Emergent topic detection

3.2 Structured Sources

As discussed in Deliverable 2.2.2, Wikirate will soon be introducing many new kinds of Sources, such as primary source material entered directly on Wikirate.org.

Another very important arena of new source material will be structured data sources, particularly once we introduce the Metric and Measurement components of the coming rating system. At this point we anticipate many extensive integrations with sources such as WikiData and CDP.

A central premise of these kinds of integration is that all structured data will be “escorted” onto the platform by a Wikirate community member. Following the pattern of a Claim creator or the creator of an individual Measurement card, these community members will be responsible for assuring that the data on Wikirate is accurately sourced and accurately represents the data of the original source.

3.3 Social Media

Wikirate has made the strategic decision *not* to focus on social media integration in its initial months in order to avoid a variety and scale of attention for which it is not yet prepared.

That said, we have taken many steps to ensure successful social media integration in the future, including:



- Limiting claims to 100 characters, which leaves 40 characters for shortened links and automated tags important to Wikirate (company, topic, etc) when tweeting a claim
- Restructuring of account handling for compatibility with social login services (via Facebook, Google, Twitter, etc)
- Use of social media feeds for finding suggested sources

4 Advancement of the state of the art

4.1 Present

Deep Atomism

Wagn's foremost organizing principle is "everything is a card"; cards are Wagn's *atoms*. While many CMS and wiki systems have some analog of the card (Drupal has *nodes*, WordPress has *posts*, etc); these other systems also have many other parallel structures. For example, if adding microstructure to a Drupal node, a developer will often create a new relational database table, which is decidedly *not* a node. This practice weakens the *atom* metaphor, because systems are then built upon building blocks that are neither atoms nor built from atoms. By contrast, when a Wagn developer wishes to add microstructure, this is accomplished by *nesting* cards within cards. We refer to this distinction as *deep atomism*.

This atomic approach is central to nearly all advances of the state of the art presented here, both at the platform and the website level. The systemic consistency it produces opens up many rich patterns of data re-use. When combined with Set-based configuration, it provides a rich framework for forming deeply reusable patterns for contributors at all levels, from Wagn core and mod developers to site administrators to users of Wikirate.org.

It's worth noting that Wagn did not *invent* deep atomism; many low-level computational systems have embraced very similar approaches. Instead, Wagn's core advances are in making atomic organizing patterns accessible to non-technical groups and individuals within an approachable web framework.

The basic atomic principles were in place before the start of the current grant, but adherence to the principle has been further enhanced in year one as pre-existing non-card structures were integrated into the card framework. Most notably *users*, which were formerly tracked in a separate (non-card) table, are now fully implemented as cards. Similarly, JavaScript handling and emails have been translated into cards in recent months.



The ensuing sections will explore some of the benefits of Wagn’s deep atomism and how it advances the state of the art.

Atomic Sets

As you would expect, an atom is rather simple. A card has a unique id, a unique (but changeable) name, typed content, and a revision history. But the atoms become powerful when we organize them into Sets and enhance the capabilities of all cards in a given Set.

A Set is a configurable group of cards and may be as broad as “all cards” or as specific as “this specific card” or somewhere in between. Sets of cards can be extended in code by “mod developers” or through the web interface by “wagneers”.

As a simple example, all permissions (create, read, update, and delete) are implemented through rules attached to sets of cards. The default read permission on Wikirate.org is, not surprisingly, “Anyone”. But this permission may be overridden with read rules on narrower sets of cards. For example, read permissions on cards that contain account information are restricted to account owners and administrators. Consistent with Wikirate’s emphasis on transparency, this is currently the only set of cards not visible to anyone (though most cards can only be updated by “Anyone Signed In”).

The patterns that emerge from Set-based configuration are powerful, and we’re just beginning to explore them. While we are using the term “Set” in the same conceptual way as “Set theory”, we have barely scratched the surface of what that body of thought offers as potential avenues of organizational creativity. We hope that our advances here will open the door to much deeper explorations.

Comprehensive revision tracking

Given that “everything is a card” and “all cards have revision histories”, one can infer that “all revisions are tracked”. At the start of the grant, this was not entirely true, because the revision history was comprised entirely of content edits, and not of changes to other core card attributes, specifically name and type. Now all core attributes are tracked. We know of no other wiki or CMS system with such comprehensive tracking.

Conceptually this means that any kind of change (an image update, a new email, a rule, etc) can be reverted. Indeed development in support of such deep reversions has begun, though current implementation has focused on single-card reversions rather than broader system rollbacks.

This tracking is also highly valuable in a transparency-driven site like Wikirate. Because they are all performed in cards, changes to site configuration, permissions, etc. may be made publicly



available to community review. Of course, the nuanced permissions system means it is also easy *not* to show such information in places where it is deemed inappropriate.

New Architectural Pattern

Wagn's emphasis on reusability extends all the way down to its architecture. Wagn is pioneering a new architectural pattern that we hope will inspire other platforms to make use of its organizing patterns.

The pattern's current working name is MoFoS for "Model – Format – Set", but we're looking for a name with a bit more gravity. The core of the architecture is this:

- one Model ("Card", of course")
- many Formats (html, txt, json, etc). All format objects are associated with a card.
- both can be subdivided into Sets

One of the most unusual qualities of this architecture is that it presents a mechanism by which code developers and site administrators can act upon similar constructs, so that they can begin collaborating immediately upon creation of a new instance. For example, formats have special methods called "views" that are used by site administrators to determine how they want to display card data. Mod developers may create and alter views in order to extend administrators' capabilities.

Another great advantage of the architecture is the simple clarity of its web and backend API's. Every action ever performed on the system is accomplished as a CRUD (Create, Read, Update, Delete) action on a card. This leads to a uniquely simple RESTful API. It also means that all data transformations are expressed as *events* on a specific set of cards, making the API both very uniform and fully adaptive to new structures created by site administrators.

Imed Hammouda, an Associate Professor at the University of Gothenburg in Sweden who studies software architectures, has approached Grass Commons to propose writing an academic paper about the new architectural pattern, and Grass Commons sent back preliminary materials in support of such an endeavor.

Sources and Claims

Wikirate adds additional structure to the standard wiki format – in the form of sources and claims. This additional structure is required because of the nature of Wikirate's subject – not all of the sources of information with respect to companies' behavior are equally reliable and in agreement with each other, these sources can be large and information-dense and the information contained therein is not of uniform importance. To take the example of a company CSR report - these are often lengthy documents, they may contain important information but



also material which could be described as ‘greenwashing’. It can be difficult for an individual to determine if a source is credible, and to determine which aspects of the source document are most important. The *meaning* of the sources which Wikirate deals with is often ambiguous and must be the subject of a process of discussion and deliberation by the community. For these reasons, it is not sufficient for a Wikirate article to reference sources directly in the same manner as a resource like Wikipedia. Wikirate brings these sources on to the website as entities or units of content which can be discussed, rated (soon) and dissected.

Claims are a tool for dissecting sources or extracting key points. Articles cite claims, which in turn are based on sources. When an individual reads an article on Wikirate they should be able to follow a link to the page of any individual claim and on this page see the community’s assessment of that claim. The claim should also link back to the source, for large sources specifying where the claim originates within the source – and showing the reader the community’s assessment of the source itself. Wikirate will provide not just a link to a source, but community-generated assessments of whether the claim is valid and whether the source is credible. Claim and Source content types and the structural relations between these and articles are already present on the Wikirate website – the tools through which the community can assess this content are in development and discussed further under the Data Quality header in the ‘Future’ section below.

Multimedia Duplicate Detection

Duplicate detection is a feature that hasn’t been integrated to a great extent on wiki-like platforms such as Wikirate because the design and deployment of an efficient image search system for duplicate detection constitutes a complex design problem calling for numerous decisions with respect to individual multimedia analysis components. Multimedia duplicate detection is mainly focused on very fast and scalable image search and one of the main challenges to be addressed within T5.1 is the development of a compact yet distinctive image representation and a scalable image indexing and retrieval scheme. To this end, the combination of VLAD aggregation and product quantization (PQ)-based indexing was selected as the main implementation basis for Wikirate. Moreover, the VLAD/PQ scheme coupled with SURF descriptors yields competing accuracy compared to state-of-the-art image search systems, while being highly scalable.

Topic Detection

Topic detection algorithms, at a high level, are split into two classes: a) document-pivot and b) feature pivot. Methods in the first class attempt to cluster together documents that refer to the same topic, whereas methods in the second class attempt to cluster together terms that represent a topic according to their co-occurrence patterns. We used a document-pivot method that utilizes an efficient hashing scheme (LSH) to rapidly find nearest neighbours of



incoming documents, as well as a number of novel feature pivot methods. These include a graph-based method that clusters terms according to their pairwise co-occurrence patterns, a method based on frequent itemset mining and a “soft” version of it, as well as BNgram, a method that selects the most bursty n-grams that appear in the collection of documents and subsequently clusters them according to co-occurrence frequencies. Moreover, we compared these approaches to a standard probabilistic topic detection algorithm, Latent Dirichlet Allocation (LDA). Out of these approaches, BNgram achieves the best performance.

4.2 Future

Multilingual Wiki

Among our most exciting advances is Wagn’s approach to multilingual content, which is scheduled for deployment at the time of Milestone III in month 18.

Common practice among wiki sites is to have completely separate sites for separate languages. There are many wiki platforms with support for instantiation in one of multiple supported languages, and a few with support for linkages between pages addressing related content in different language, but no current system supports rich multilingual interactivity.

The core of our plan is to support context-appropriate translation rules on a granular level and to engage users to help in direct translation of the most critical content on the site. Users will be able to convey what languages they understand, and Wagn/Wikirate will only present them content in these languages based on their order of preference.

Multilingual support on Wikirate will involve socio-technical innovation. The task of translating a piece of content from one language to another will be presented to users as a possible means of contribution alongside other core actions like article editing and claim creation. This task will also be included within Wikirate’s incentives and gamification scheme. The goal is to represent translation of a piece of content from one language to another as an action with the same level of desirability as the creation of that content – i.e. a user who translates a claim into a new language has contributed ‘as much’ as the user who submitted that claim to Wikirate in the first place. Users will input their language knowledge/preferences to obtain the best possible Wikirate experience – and this information will also allow us to present appropriate translation requests to users unobtrusively on the site.

This is possible to tackle in a granular way on Wagn because content is organized into cards. Sets of cards will follow different translation patterns. Some, such as the cards holding numeric



values like company measurements or vote counts, will be universal. Some, like claims, will be directly translated from one language to another. Others, like discussion cards, will be independent in different languages.

Details of the proposed functionality are discussed in more depth in deliverable 2.2.2.

Ratings Marketplace

Many projects have proposed mechanisms for rating facets of corporate social responsibility. Different projects have succeeded in various ways, however no one project has presented a high quality, accessible, transparent and digestible mechanism; which responds to the fact that individuals may have differing value profiles.

There have been many previous attempts of projects attempting to rate companies. However there has not been a successful attempt, which satisfied all of the above criteria. Many projects satisfy some of the criteria, e.g Goodguide, which is highly accessible/ digestible; however the numbers are driven by 'expert' research. Another example is CSR Hub, which has great interface, and sliding scales to represent various value profiles, however when it comes to see the sources driving ratings many are behind paywalls and it's difficult to see any correlation between the content and the ratings output. Other websites, such as Brandkarma, are highly digestible for a user, but by looking to score companies based on customer experiences and sentiments, the ratings are highly subjective.

Identified Problems to solve:

- Value dynamics are different for different people (value profile) subjectivity of ratings
- Information driving ratings and metrics needs to be **transparent** and **accessible**
- Ratings themselves need to be accessible to wide stakeholder group (digestible), but not at the expense of quality
- Information gaps that drive ratings should be identified as all data will not necessarily be publicly available
- There should be positive and productive interface with companies, with transparency being rewarded

Our core innovation in this arena is that we're assembling not so much a ratings mechanism as a *ratings marketplace*. On Wikirate users will be able to create their own corporate metrics, and the best metrics can percolate to the top, while the poorest metrics effectively disappear from sight. The result is an adaptive system that can fluidly respond to new research, new data, new technologies, and new ideas.



Perhaps even more important is the fact that we've designed this system not only to assemble and distill existing data but also to incentivise companies to make new data available. The primary mechanism for this is the *WRIT (Wikirate Index of Transparency) rating*. A company's WRIT rating is a measure of how well a company has made information available to the Wikirate community. It will be the most prominent rating on Wikirate.org and the only one endorsed by Wikirate itself. Via this rating, companies will be rewarded for responding to inquiries and punished for failing to do so. The rating is weighted based on metric importance. So that if the community has determined that CO2 emissions are an important metric, the availability of this information will have a significant impact on a company's WRIT rating.

This design advances the state of the art in two keys ways:

- It will be the most transparent CSR ratings system ever implemented, including deep support for discussion, refinement, and explication of every metric involved

It is more responsive than any ratings system proposed; it reacts fluidly to data availability, corporate responsiveness, and citizens' perception of the relative importance of different metrics

Data quality mechanisms

Wikirate aims to provide a set of tools which enable users to collectively assess the information which is available in relation to companies' behavior – **and** structure these tools in such a way that they combine as a process for the production of the resource. This advances the state of the art by making the process of wiki-type production much more granular and accessible – enabling it to be applied to the difficult task of *making companies clearer*.

The data quality mechanisms which are in development apply a form of up/down voting to the assessment of **claim importance** and **source credibility**. Up/down voting systems are not novel in themselves, but Wikirate's attempt to integrate these deeply within a production process represents an advancement of the state of the art.

Voting scores, and particularly the pages where these scores are used to determine which items are shown, will offer a clear indicator of community consensus on the quality of content. When a user has spent some time reading/browsing content on Wikirate they should have a clear sense of the types of content which the community values most highly – this will stand them in good stead should they decide to contribute content themselves, it will also serve as a way in which the Wikirate community's identity can be defined.

Voting on Wikirate will reflect an item's quality – and items which score more highly will be prioritized for usage in the 'next step' of the production process. Sources will be evaluated on



‘credibility’, and more credible sources will be offered more readily for users’ consideration in making claims.

Claims will be evaluated on ‘Importance’ – and articles should try to cite the most important claims which are relevant to the company and topic. A list of ‘Important’ claims (and later ‘Recent Important’ claims) will double as a kind of news feed on company behavior – this could encourage repeat visits from users even when they aren’t returning with an intention to contribute content. Such lists will also serve as useful resources for users who want to write articles, as they will contain important claims and visual indicators of whether those claims have been cited yet. Within articles the claims will be colour coded so that it is possible for a reader to determine at a glance whether the article cites important claims or not. This is of particular importance for Wikirate because in assessing a company’s behavior it is often important to consider not only the information which is presented but information which may have been presented but is not. A Wikirate article could be manipulated to paint a more positive picture of a company by excluding the most critical claims about that company from the article or marginalizing them. When the community has decided (through voting) that a claim is important - the colour-coding of claim citations within articles, and the prominent display of the most important claims which haven’t been cited beneath the article, will make it much more difficult to exclude an important claim from the article.

In contrast to a site like Reddit, claims that are rated as important on Wikirate will be ever-present – they will not slip out of view after a day or two. This means that it is of greater importance that Wikirate’s user votes produce accurate, reliable scores. To this end, user voting on Wikirate will be a matter of public record – it will be possible to see which users have voted on which items and in which direction. This is a departure from how other sites with user voting work, where anonymity of votes is the norm. While Wikirate’s voting cannot be considered an advancement of the state of the art in a technical sense – structuring the site in a way that individual users and the community as a collective must stand by their votes (or correct these when they are shown to be errant) could certainly be considered innovative in a socio-technical sense. On a site like Reddit user votes and the decisions they result in have a disposable quality – if Wikirate can ensure that voting scores always reflect current community consensus on an item’s quality this would be a significant advancement in terms of the expression of collective intelligence through user voting.

Emerging Topic Detection

Collecting trends and insights from public services can be useful to detect new topics that are not tracked by Wikirate (WP5). There are several platforms that offer access to trends from social media. One of them is Twitter itself through the discover tab available for each twitter user. Other tools that show twitter trends are TweetDeck and Topsy . Google Trends and



Google Insights also provide information on trending and hot searches. An alternative source of trending information though can be obtained by periodically monitoring the most popular social multimedia, e.g. YouTube videos and Flickr photos. Through the APIs of multimedia sharing platforms, it is possible to obtain lists of most popular media items per topic. This can in turn be used for extracting trending keywords (e.g. by processing the titles and other textual metadata of the trending content items).

Duplicate and Topic detection

Plans for development of the duplicate, topic, and emerging topic detection systems outlined above are as follows:

- Taking relevance feedback from community members into account to further enhance the source recommendations
- Personalized recommendations based on user feedback
- Optimization of the algorithms in terms of accuracy
- Integration of other WVP5 features in Wikirate.org
- Concerning multilingualism, separate source collections should be set up for each supported language
- Improve technical documentation (mostly API related)

5 Recommendations

Most of our recommendations for technical improvement are included inline in the sections above. For example, the Scalability section goes into great detail about many routes for future improvement, and the Testing section outlines many specific needs.

It's also true that all the plans discussed in deliverable 2.2.2 reflect broad contributions from members across the Wikirate Consortium who participated in design discussions on our internal Wagn site, in conference calls, over email, and in Huddles.

That said, we have reserved some additional general recommendations for this section because they don't clearly fit elsewhere. Specifically:

- I. Greater clarification of which aspects of the Wikirate.org site should be subject to our design process (which is managed with many tools outside of Wikirate.org), as opposed to community process conducted entirely on Wikirate.org itself. There is currently a rough understanding that structured pages will require design attention and their initial implementation is sped by the use of external development tools, but the boundaries of such designed pages are not entirely clear.



2. Deeper engagement of the Wikirate community in the development process. As our community grows, we should make a concentrated effort to reach out for their feedback, both for the value of their input and to deepen their investment in the project.
3. Enhanced support for and prominence of discussion cards on Wikirate in order to support both of the above.
4. Improved documentation of all Wagn and Wikirate code to entice greater contribution to the open-source code bases.