High Quality Information Delivery: Demonstrating the Web in Your Pocket for Cineast Tourists

Florian Stahl*, Adrian Godde†, Bastian Hagedorn†, Bastian Köpcke†,
Martin Rehberger†, Gottfried Vossen* † ‡
∗ WWU Münster, ERCIS, Florian.Stahl@ercis.de
† WWU Münster, CS Dept., {a.godde | b.hagedorn | basti.k | m.rehberger}@wwu.de
‡University of Waikato Management School, vossen@waikato.ac.nz

Abstract: The Web in Your Pocket is an information platform providing users with high quality content, sourced from the Web and curated by human experts. We describe the process as well as its implementation and elaborate on our demo case: tourists wishing to explore remote filming locations.

1 Introduction

Online information has exploded over the last decades. This implies that it becomes increasingly harder to find information – in the sheer mass of information available – that is really relevant to a particular situation: the proverbial needle in a haystack. This may be because of complex queries requiring data integration [Cer10] or because comprehensive semantic knowledge is necessary, yet not available [HSBW13]. Furthermore, keeping information available when being offline is currently not integrated in search processes. In some cases, this can enhance the perceived utility of the system tremendously.

To overcome this dilemma, we have developed a framework that does not answer queries solely based on a pre-assembled index, but based on a subject-specific database that is sourced from the Web, integrated and curated by domain experts, kept up to date automatically, and dynamically generated based on vast user input. Since our framework is supposed to make relevant information accessible on a mobile device when no Internet connection is available, we refer to it as Web in your Pocket (WiPo) [DSV12]. The contribution of this demo lies in presenting the implementation of a complex data gathering, processing, transforming, editing and curating process that enables users to find relevant information which can be persisted offline.

Imagine a tourist travelling New Zealand who wishes to visit locations where the Lord of the Rings films have been shot. Confessedly, Web sites such as *http://www.movie-locations.com/* provide information on filming locations, but they do so in an un-integrated way. Applying the generic WiPo framework to the situation of cineast tourists, we realised a platform on which information about filming locations is presented in a way that incorporates information about the film itself (e. g., sourced from Wikipedia), images, snippets, or film trailers, (e. g., sourced from Youtube), and general tourism information (e. g., sourced

from local tourism and accommodation Web sites). Furthermore, we provide a means to make this information available offline, so that tourists do have them at hand, when they are for example in a New Zealand region where no Internet connection is available. We mention that the system supports a number of business models. For the demo case a free-to-use model is most appropriate. The operating cost could either be covered in a Wikipedia-like fashion where people contribute time and money for the sake of the greater good. Alternatively, a third party with a severe interest in publishing the information, such as local tourism agencies, could step in as a sponsor.

To our knowledge, there are only a few approaches that combine curation and Web Information Retrieval. Sanderson et al. [SHL06] suggested to apply a procedure where predefined queries are sent to pre-registered services on a nightly basis. The retrieved information is audited by curators who decide upon its relevance. A similar harvest and curate approach was suggested by Lee et al. [LMH09] as ContextMiner (`http://contextminer.org/`). However, neither of these is able to make information available off-line on a mobile device or offers advanced features such as highly-customisable sources or curation as a service, i. e., curation done *for* the user rather than *by*. This makes WiPo a new and innovative approach to information gathering. In this paper, we will first recapitulate the basics of the innovative WiPo approach and the prototypical implementation which serves as proof of concept and as a vehicle for practical experiments. Subsequently, we demonstrate WiPo's capabilities in the area of film tourism.

2 The WiPo Concept & Architecture

WiPo is a process-based approach to information gathering, Web search, and data curation shown as a high-level Petri Net in Figure 1. It is generic in nature and can be applied to any domain, given domain specific knowledge. The implemented architecture follows the client server principle based on a REST API. Currently, the client is a browser-based GUI, but could be an arbitrary application that can communicate with the server's REST-based Web interface. A comprehensive description of the implementation can be found in [SGH+14].

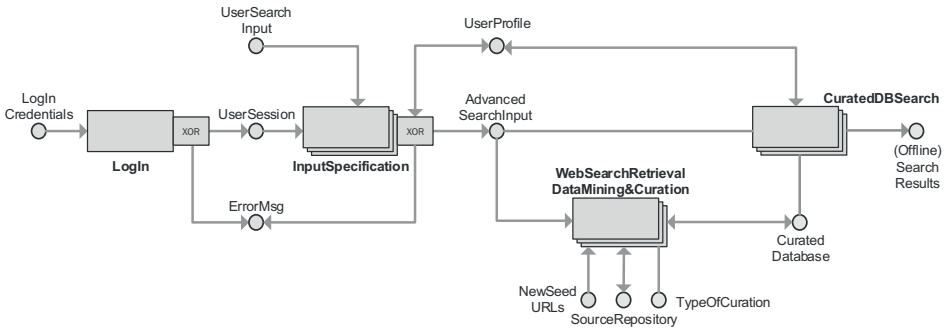

Figure 1: The Overall WiPo Process

First, users log on to the system, resulting in a *UserSession*, and provide *UserSearchInput* which – if valid – can either be profile data that is populated to the stored *UserProfile*, or *AdvancedSearchInput* which must contain a search query and can be enhanced with URLs or private files. *InputSpecification* checks this and creates *AdvancedSearchInput*. This object is then used in two ways. First, it is used to perform *WebSearch, Retrieval, DataMining, & Curation*. In this step, appropriate sources are selected from a *SourceRepository*, to which curators can supply *NewSeedURLs*, and a focused crawl is performed on the selected sources as well as on user-supplied URLs contained in the *AdvancedSearchInput* object. The crawl results undergo data mining and eventually curation (i. e., selection, editing and integration). Eventually, high-quality results will be stored in the *CuratedDatabase*. Secondly, search keywords are read from *AdvancedSearchInput* and are used to query the *CuratedDatabase* in *CuratedDBSearch*. Last, relevant results – considering the user profile – are delivered to the user who then may chose to which to persist offline.

In the prototype, selected URLs are crawled by an Apache Nutch instance. After successful crawling, meta data, i. e., fetch time, last modified time, etc. are determined (stored in the source repository) and content is extracted and passed on to curation – WiPo's distinguishing feature, intended to assure high quality information. In order to achieve this, we suppose that an expert will verify or modify the data mining results prior to their delivery to the user. However, in some cases this expert could also be an algorithm or an externally provisioned service or even a crowd (referred to as *TypeOfCuration*). The curation result will be documented in an indexed repository – the *CuratedDatabase* – implemented using mongoDB and Apache Solr. Thus, WiPo differs from other Web retrieval approaches in that it does not try to solve everything automatically but relies on human curation. A first step towards more automation – since curation is a laborious task – is to allow for automatic propagation of changes to an original source. Therefore, curators have the option to choose what happens, when the source of a curated document changes. The options are *keep*, i. e., text in the curated database remains as is, even if the underlying source changes, *notify*, i. e., curators will see the document again, once it has changed and can then decide what to do, and *auto*, i. e., changes will be propagated unchecked.

The first prototypical implementation which builds the basis for this demo has some minor restrictions. First, it does not yet consider user-supplied documents and as a consequence *InputSpecification, WebSearch, Retrieval,* and *DataMining* are limited. Furthermore, curation is to date restricted to manual curation. Nevertheless this implementation is caplebale of demonstrating the whole process as will be evident in the next section.

3 Use Case & Demo

In this demo[1], we will demonstrate how curators maintain the curated database, as well as how users search for information and how they can tailor the search towards their needs. Therefore, we will present a) the usual work of and b) the data flow between curators and users by the example of cineast tourists.

[1] A video demo can be found at *http://dbis-group.uni-muenster.de/y/wipodemo*.

In the curation demonstration, we will demonstrate the creation of a curated document by New Zealand (NZ) tourism experts. In this context s/he will highlight the use of text segments to combine different sources, multi media content such as images or videos, and map data. As an example, s/he will create a document featuring the Hobbiton movie set containing information regarding, opening hours and entrance fees (from the original Web site) as well as what to particularly look for (sourced from own knowledge or blog entries). This information is enriched with images and video footage as well as general tourism information for the Matamata region (sourced from a variety of Web sources).

Subsequently, the tourist form the introduction with a special interest in Lord of the Rings will search for information on filming locations in New Zealand. This will return for instance the newly created document on the Hobbiton movie set including all additional information such as what else to see in the Matamata region. Given that the tourist knows about the poor mobile Internet coverage in rural New Zealand areas, they wish to persist the information for offline usage with the aid of Evernote. It has to be pointed out that Evernote was a choice of convenience and any other "read-it-later" tool providing an API could have been used. Further, we will go into detail about factors influencing the search result ranking such as previous rating of received documents and the explicit user profiles. To this end, our user will for example change their preference from the North Island to the South Island. As a result, Hobbiton will disappear and new locations will be presented such as the Dry Creek Quarry near Wellington which served as location for Minas Tirith and Helm's Deep.

Following the demonstration of the search using the curated database, we will show how URLs provided by users (e. g. information on a film location tour not yet known to the system such as *http://www.lordoftheringstours.co.nz* are processed. This includes crawling of the supplied URLs, the subsequent curation of the results as well as how the results are eventually presented to users.

References

[Cer10] Sefano Ceri. Chapter 1: Serach Computing. In Sefano Ceri and Marco Brambilla, editors, *Search Computing*, volume 5950 of *Lecture Notes in Computer Science*, pages 3–10. Springer-Verlag Berlin Heidelberg, Berlin and Heidelberg, 2010.

[DSV12] Stuart Dillon, Florian Stahl, and Gottfried Vossen. Towards The Web in Your Pocket: Curated Data as a Service. In *Advanced Methods for Computational Intelligence*, pages 25–34. Springer-Verlag, 2012.

[HSBW13] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence*, 194(0):28 – 61, 2013.

[LMH09] C. A. Lee, R. Marciano, and C.-Y. Hou. From harvesting to cultivating. In *Proceedings of the 9th ACM/IEEE-CS joint Conference on Digital Libraries*, page 423, 2009.

[SGH+14] F. Stahl, A. Godde, B. Hagedorn, B. Köpcke, M. Rehberger, and G Vossen. Implementing the WiPo Architecture. In *Proceedings of the EC Web*, München, 2014.

[SHL06] R. Sanderson, J. Harrison, and C. Llewellyn. A curated harvesting approach to establishing a multi-protocol online subject portal: Opening information horizons. In *6th ACM/IEEE-CS Joint Conference on Digital Libraries*, 2006.