RENDER
FP7-ICT-2009-5
Contract no.: 257790
www.render-project.eu

RENDER

Deliverable D1.1.2

Algorithms and infrastructure for time-contextual access to wiki-like items

Editor:	Denny Vrandecic, KIT
Author(s):	Denny Vrandecic, KIT; Angelika Adam, Wikimedia; Gerrit Holz, Wikimedia
Deliverable Nature:	Report (R)
Dissemination Level: (Confidentiality)	Public (PU)
Contractual Delivery Date:	September 2011
Actual Delivery Date:	September 2011
Suggested Readers:	Researchers working with historical data from Wikipedia, NLP researchers on trend detection
Version:	1.1
Keywords:	NLP, Wikipedia, time, temporal, Corpex

Disclaimer

Full Project Title:	RENDER – Reflecting Knowledge Diversity
Short Project Title:	RENDER
Number and Title of Work package:	WP1 Data collection and management
Document Title:	D1.1.2 – Algorithms and infrastructure for time-contextual access to wiki-like items
Editor (Name, Affiliation)	Denny Vrandecic, KIT
Work package Leader (Name, affiliation)	Maurice Grinberg, Ontotext

Executive Summary

This deliverable explains the need for time-contextual access to wiki-like items and how to achieve it. It further shows an analysis of such a time-contextual corpus.

The need is given by two use cases: 1) readers who would like to see how Wikipedia (or another wiki) looked like in the past, and 2) researchers who need to be able to reproduce a previous state of the content for further experiments, e.g. for NLP corpus research.

We present a survey of three existing solutions: the Internet Archive, Memento, and JWPL. We show that the Internet Archive can deal with the first use case, and JWPL with the second. This fulfils the requirements of the deliverable.

In addition, we extend our previous corpora analysis to recreate corpora for past years, and offer a first glance into the corpora. The creation of the corpora took longer than expected, and thus this deliverable only reports on first results. We will provide a full dataset once it is processed.

Table of Contents

List of Figures

List of Tables

Abbreviations

API	Application Programming Interface
Corpex	CORPora EXplorer
GPL	GNU Public License
HTML	HyperText Markup Language
HTTP	HyperText Transport Protocol
JWPL	Java WikiPedia Library
NLP	Natural Language Processing
SMW	Semantic MediaWiki
XML	eXtensible Markup Language

Definitions

Complete dump	A file containing all revisions of all pages in Wikipedia
Corpex	Website to explore NLP corpora, with quick and easy access to frequency data, both provided as visualizations and through a RESTful API
Corpus	A set of text used for NLP
Current dump	A file containing the current revision of all pages in Wikipedia
Revision	The content of a Wikipedia page after one edit
Term	Also: Word form, a distinct word that occurs in a corpus
Wiki	Website that allows the simple editing of content
Wikipedia	A project for creating a free encyclopaedia; refers also to the encyclopaedia itself
Word	The occurrence of a term

1 Introduction

This deliverable describes algorithms and an infrastructure for the time-contextual access for wiki-like items. This section will discuss the motivation for providing this work and how it fits into the RENDER project [1]. The second section is a survey of existing tools and an evaluation if the existing tools meet our needs. The third section describes the extension of the Corpex corpus developed in RENDER with time-contextuality, thus validating the motivation for this work. It also offers some preliminary results, but it has to be understood that the experiments are still ongoing. Section 4 concludes the deliverable.

1.1 Wikipedia as a language corpus

Wikipedia is a free encyclopaedia hosted by the Wikimedia Foundation and edited by anyone. Wikipedia is provided in more than 250 language editions, and for many of these language editions Wikipedia is the first encyclopaedia ever. In most language editions, Wikipedia is the biggest encyclopaedia available in the given language.

Besides being an encyclopaedia, the collaborative development and the openness of the Wikipedia project create a huge number of interesting data sources that are still widely underexploited. Every single edit in Wikipedia is archived over its more than ten year history, the collaboratively created history of the Wikipedia articles can be analyzed in order to understand the development of collaborative artefacts, but also to understand collaboration in general. At the same time also the software that Wikipedia is running on is completely open source, in particular the wiki engine MediaWiki, which was originally developed to support Wikipedia and later turned into a general purpose open source wiki. The development of the software can be analyzed, the discussion around its development is archived in mailing lists, and every single code commit is archived.

The German chapter of Wikipedia participates in the RENDER project. One of the goals of the RENDER project is to provide easier access to data surrounding the Wikipedia project. One of these datasets was provided in Deliverable 1.1.1, by creating frequency distributions over all the terms in Wikipedia for dozen of languages, thus providing novel language corpora that can be used for NLP, as well as a novel user interface to access the data via the Web [2].

Websites:

- Corpex: http://km.aifb.kit.edu/sites/corpex

1.1 The past of Wikipedia

Even though Wikipedia retains a record of every single edit, it is far from trivial to recreate a view of a page at a certain point in time in the past. In this section we will comprehensively list all issues that appear. Note that none of the solutions discussed in this paper solves all the issue involved in allowing the easy access to Wikipedia's past. We discuss how the solutions solve the issues, and discuss for which tasks a given solution is sufficient.

What is the current situation?

A wiki like Wikipedia consists of a set of pages or articles. Each page consists of a time sequence of revisions. In the database, each revision is saved completely, not as the difference between two revisions. So for each page it is trivial to figure out the revision that was current at a specific point in time. A tool can easily retrieve the revision that was current at a specific point of time and render that revision. MediaWiki itself actually offers that feature, so the user can browse previous revisions and get acquainted with the history of a specific article.

There are a number of problems though, that the MediaWiki system ignores:

- MediaWiki can use page inclusion and templates. In the wiki text of a page, a template or other page is being called, which will be included at that location and rendered. These template calls might be parameterized. MediaWiki always resolves a template call with the most current version of the template instead of the version that was current at the time of viewing the page. Also, the history of an article only lists the revisions of the given article, it does not consider revisions of the included templates. This means that the history of an article does not even show when the appearance of the article has changed, based also on called templates, but only when the wikitext of the given page has been modified. Finally, the MediaWiki API is not based on absolute points in time but rather in discrete revisions with timestamps – requesting the version of a page at a given point in time always requires to first query for the correct revision id.
- MediaWiki has a number of magic words or function calls that tell the parser to execute certain functions and replace them with the result of these functions. Among those are a number of functions that result in different outputs depending on a number of circumstances, e.g. magic words to display the number of pages in the wiki, the number of pages in a specific namespace, the number of registered editors, the date, the day of the week, etc. Additionally these function calls can also be conditional, and thus it would be possible, although not so useful on an encyclopaedia page like Wikipedia, to display different texts depending on the day of the week, the month, etc. This is used, for example to display the "This day in the past" section on the Main Page of Wikipedia.
- Semantic MediaWiki (SMW) adds a similar layer of problems. Semantic MediaWiki's *ask* parser function call executes a query against the current knowledge base – no matter if we are viewing an old version of the page or not. In order to enable to query previous versions these would either need to be saved, or recreated on the fly – the first one is unfeasible due to space constraints, the second one due to time constraints. The interplay of SMW and parser functions furthermore makes the system Turing complete, resulting in the following problem where the value is undefined:

  ```
  [[value::{{#expr:{{#show:{{PAGENAME}}|?value}}+1}}]]
  ```

  This means we give the property "value" the value of itself plus 1 (Pagename is resolved to the name of the current page, itself. Thus the show query asks for the property value of this page. (In comparison to the *show* parser function the ask parser function returns besides the value of the property also the name of the property. This is undesired since we store the result of that parser function call as a property value.) The *expr* function calculates the mathematical expression of the result of the ask query, i.e. the value of the page, plus one. Then this result is assigned to the value of the page). Whenever the parser is being executed, the value of the property will change. If the parser is executed, though, is an intricate question resolved by the interplay of the caching settings, accesses to the page, accesses to related pages, and other factors. Thus it makes it basically impossible to recreate the value of "value" for a given point in time.
- The local settings of a wiki, as well as further settings in the database (like the list of interwiki links) also play a decisive role how a page is being parsed, display, and rendered. In order to recreate the old versions it would be required to keep complete change logs of all these settings. In most systems the local settings though are not under version control and thus cannot be recreated.
- The software version of MediaWiki itself might change. This will, from time to time, introduce new features, deprecate old ones, and thus lead to differences in the output of the program. The early versions of the wiki running Wikipedia did not have free links, for example, but instead camel cased words were used to indicate links (i.e. words that had inner letters capitalized, like NewYork or EuropE). Later, this functionality was dropped, so those links became normal words. Instead free links were introduced, i.e. anything could be a link as long as it is enclosed by double square brackets.

Some of these problems would be only solvable through a complete recreation of all steps that lead to the given situation. But complete logs of these steps do not necessarily exist. The result of this overview is the understanding that it is not possible to recreate a wiki page by the server and its knowledge.

1.2 Why history is important

There are two major use cases for accessing a previous version of an article or of the Wikipedia as a whole.

Use case 1: The first one is to satisfy the curiosity of the user in order to understand the development of a page over time. Today this curiosity is partially solved with the normal "View history" action of MediaWiki. The following figure demonstrates this.

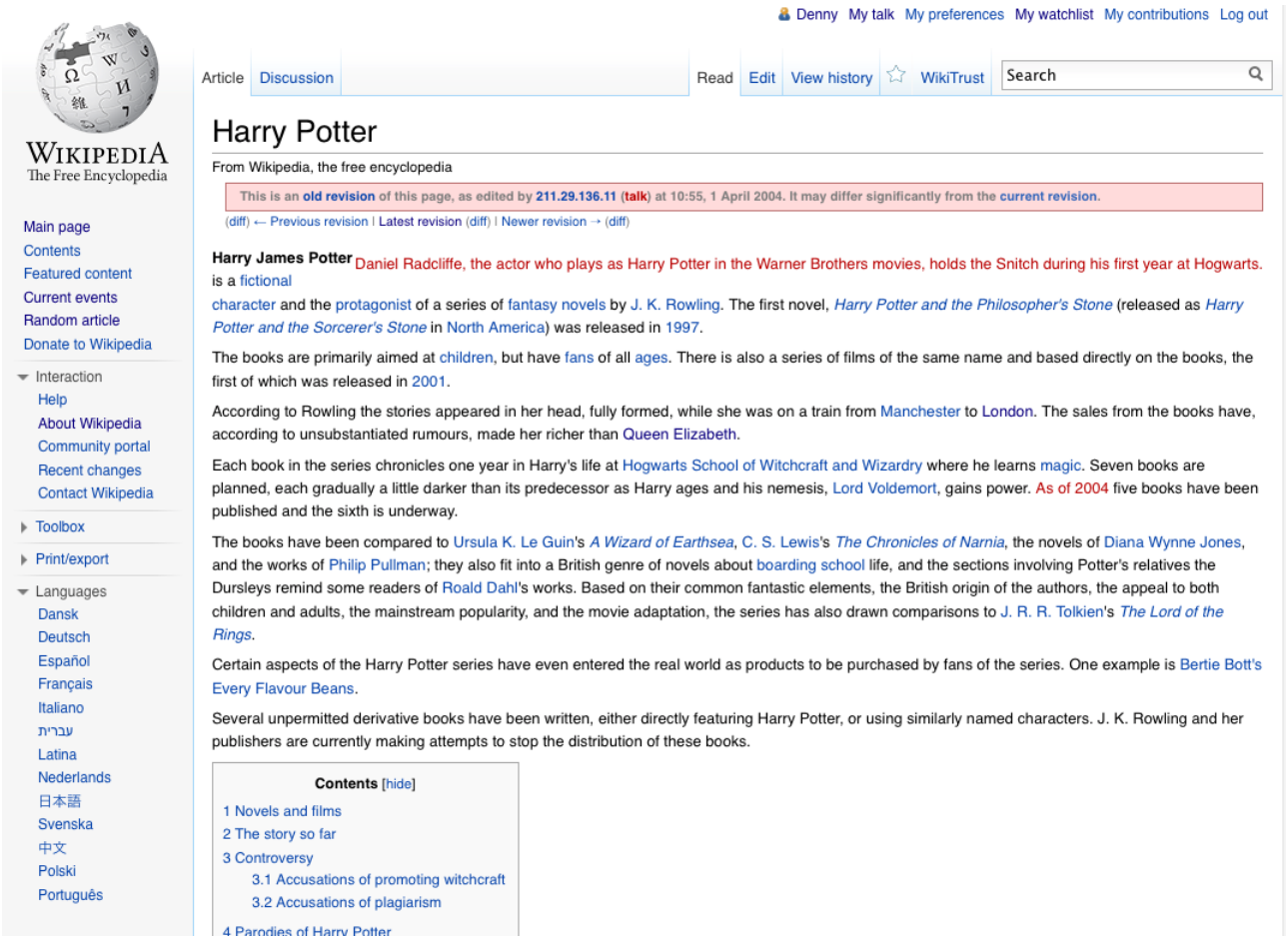


Figure 1: The Harry Potter article on Wikipedia in its revision from April 1, 2004, as seen in 2011.

What we also see is that this page uses the new logo and layout of the page. The image is also not accessible. All used templates are the current ones – and in many cases the template does not exist anymore, leaving a red link.

Even though these drawbacks all exist, we have to recognize that this is not a pressing issue. Among the many requests to Wikimedia, the request to have a more authentic view of every single previous version of Wikipedia has not been formulated loudly. The internal solution with the "View history" tab is sufficient for many users, and for some other use cases also easily accessible solutions exist (see Section 2.1).

Use case 2: The second use case is for purposes of research based on the text corpus of Wikipedia. Whereas such research is getting increasingly popular (e.g. our own work on Corpex), the analysis of the complete wiki dump happens more and more often. Most of the time though the papers provide insufficient detail on which exact version of the corpus has been used. In order to understand this problem we first need to describe how Wikipedia dumps are created.

There are two major type of dumps of Wikipedia's content. The one is a dump of the current revisions of all Wikipedia pages (current dump). The second is a dump of all revisions of all Wikipedia pages (complete dump). The complete dump is huge (measured in Terabytes) whereas the current dump is merely very big (measured in Gigabytes). Most researchers use the current dump.

Dumps are created in rather irregular intervals, and if they fail they are usually just skipped (i.e. one can download a previous dump instead). The dumps are dated, and most research papers state the date of the current dump they were using for their analysis. This is sufficient to find the correct current dump as long as old dumps are being offered by the Wikimedia Foundation. But note that these dumps require a lot of space and are redundant since they can be recreated from the complete dump anyway. Any previous complete dump is redundant compared to a later complete dump.

In short: in order to reproduce previous research results it is necessary to reproduce the previous data the research was performed on. For this it is required to recreate these then current dumps. Section 2.3 shows a solution for this task. Section 3 shows first analyses performed over this solution.

Website:

- Wikipedia dump downloads: http://download.wikimedia.org

2 Survey of solutions

We have made a survey of existing solutions that tackle the issues presented in Section 1. We found three possible solutions that tackle most of the required use cases. We present the discovered solutions in this section and discuss their advantages and drawbacks. From our use cases we recognize no need to create a new solution.

2.1 Internet Archive Wayback Machine

The Internet Archive is a non-profit organisation founded by Brewster Kahle with the mission to provide "universal access to all knowledge". One of its most prominent services is the Wayback Machine, which allows access to previous versions of a website.

In 1996 Brewster Kahle et al. wrote an early crawler that collected Web sites and saved the snapshots. The snapshots were created in irregular frequency. In 2009, the Wayback Machine was growing at a rate of about 100 Terabyte a month and had an estimated size of three Petabyte.

The Wayback Machine offers a convenient web access to snapshots of the websites at previous points in time. Users can not choose points in time where the machine did not make a snapshot, and especially early versions of Wikipedia did not have many snapshots (the earliest snapshot of the main Wikipedia pages is from July 2002, one and a half years after the start of the project).

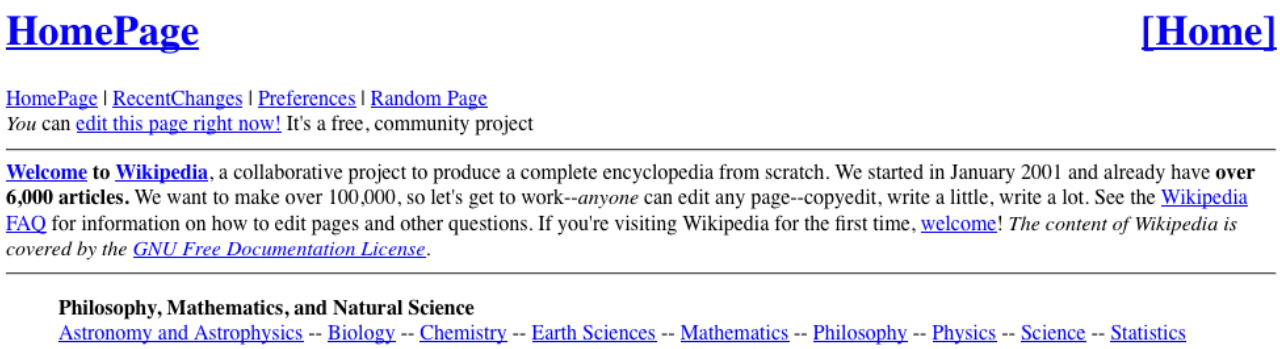
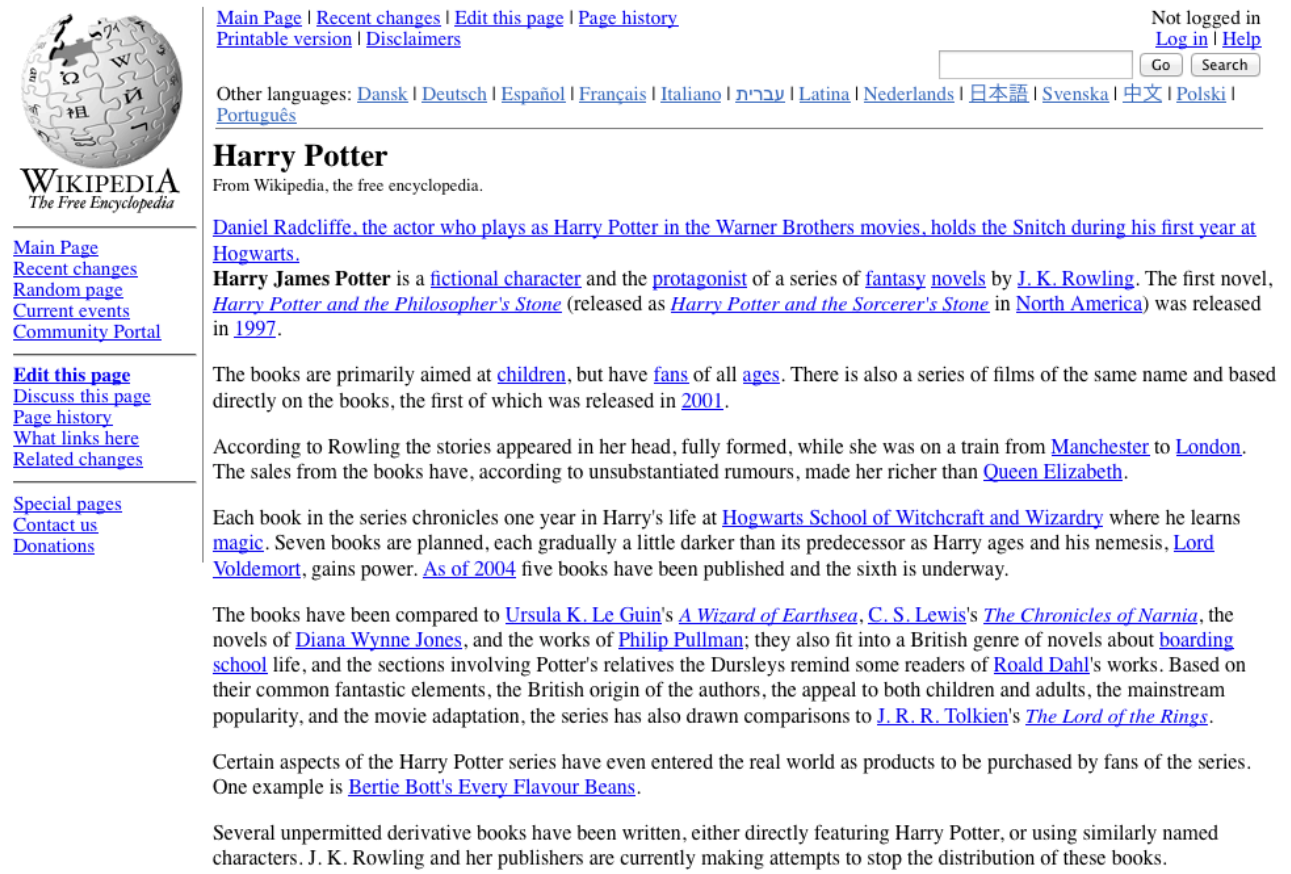


Figure 2: The Wikipedia home page from July 2001, as shown by Internet Archive's Wayback Machine.

Due to the arbitrary incompleteness of the Wayback Machine it is not sufficient for detailed analyses of the history of an article. But in order to satisfy the user in answering the question of how Wikipedia looked half a decade ago it does a better job than Wikipedia itself. The following Figure displays the Wayback Machine's rendering of the Harry Potter article of April 2004. Compare this to the screenshot in Section 1.



Main Page | Recent changes | Edit this page | Page history
Printable version | Disclaimers

Not logged in
Log in | Help
Go Search

Other languages: Dansk | Deutsch | Español | Français | Italiano | עברית | Latina | Nederlands | 日本語 | Svenska | 中文 | Polski | Português

WIKIPEDIA
The Free Encyclopedia

Main Page
Recent changes
Random page
Current events
Community Portal

Edit this page
Discuss this page
Page history
What links here
Related changes

Special pages
Contact us
Donations

Harry Potter

From Wikipedia, the free encyclopedia.

Daniel Radcliffe, the actor who plays as Harry Potter in the Warner Brothers movies, holds the Snitch during his first year at Hogwarts.

Harry James Potter is a fictional character and the protagonist of a series of fantasy novels by J. K. Rowling. The first novel, *Harry Potter and the Philosopher's Stone* (released as *Harry Potter and the Sorcerer's Stone* in North America) was released in 1997.

The books are primarily aimed at children, but have fans of all ages. There is also a series of films of the same name and based directly on the books, the first of which was released in 2001.

According to Rowling the stories appeared in her head, fully formed, while she was on a train from Manchester to London. The sales from the books have, according to unsubstantiated rumours, made her richer than Queen Elizabeth.

Each book in the series chronicles one year in Harry's life at Hogwarts School of Witchcraft and Wizardry where he learns magic. Seven books are planned, each gradually a little darker than its predecessor as Harry ages and his nemesis, Lord Voldemort, gains power. As of 2004 five books have been published and the sixth is underway.

The books have been compared to Ursula K. Le Guin's *A Wizard of Earthsea*, C. S. Lewis's *The Chronicles of Narnia*, the novels of Diana Wynne Jones, and the works of Philip Pullman; they also fit into a British genre of novels about boarding school life, and the sections involving Potter's relatives the Dursleys remind some readers of Roald Dahl's works. Based on their common fantastic elements, the British origin of the authors, the appeal to both children and adults, the mainstream popularity, and the movie adaptation, the series has also drawn comparisons to J. R. R. Tolkien's *The Lord of the Rings*.

Certain aspects of the Harry Potter series have even entered the real world as products to be purchased by fans of the series. One example is Bertie Bott's Every Flavour Beans.

Several unpermitted derivative books have been written, either directly featuring Harry Potter, or using similarly named characters. J. K. Rowling and her publishers are currently making attempts to stop the distribution of these books.

Figure 3: The Harry Potter article of April 1st, 2004, as shown by the Wayback Machine.

Denny My talk My preferences My watchlist My contributions Log out

Article Discussion Read Edit View history WikiTrust Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

Toolbox

Print/export

Languages
Afrikaans
Alemannisch
አማርኛ
العربية
Armãneashce
Asturianu
Azərbaycanca
বাংলা
Bân-lâm-gú

Harry Potter

From Wikipedia, the free encyclopedia

This article is about the fantasy book series. For the titular character, see Harry Potter (character). For the film series, see Harry Potter (film series). For related topics, see List of Harry Potter related topics. For other uses, see Harry Potter (disambiguation).

Harry Potter is a series of seven fantasy novels written by the British author J. K. Rowling. The books chronicle the adventures of the adolescent wizard Harry Potter and his best friends Ron Weasley and Hermione Granger, all of whom are students at Hogwarts School of Witchcraft and Wizardry. The main story arc concerns Harry's quest to overcome the evil dark wizard Lord Voldemort, whose aim is to subjugate non-magical people, conquer the wizarding world, and destroy all those who stand in his way, especially Harry Potter.

Since the 30 June 1997 release of the first novel *Harry Potter and the Philosopher's Stone*, the books have gained immense popularity, critical acclaim and commercial success worldwide.[2] The series has also had some share of criticism, including concern for the increasingly dark tone. As of June 2011, the book series has sold about 450 million copies and has been translated into 67 languages,[3][4] and the last four books consecutively set records as the fastest-selling books in history.

A series of many genres, including fantasy and coming of age (with elements of mystery, thriller, and romance), it has many cultural meanings and references.[5][6][7][8] According to Rowling, the main theme is death,[9] although it is primarily considered to be a work of children's literature. There are also many other themes in the series, such as love and prejudice.[10]

The initial major publishers of the books were Bloomsbury in the United Kingdom and Scholastic Press in the United States; the books have since been published by many publishers worldwide. In October 2011, the series will be released in various ebook formats through "Pottermore".[11] The books have been made into an eight-part film series by Warner Bros. Pictures, with the seventh book split into two parts; it is the highest grossing film series of all time. The series also originated much tie-in merchandise, making the

Harry Potter

The coat of arms of Hogwarts, representing the four Houses (*clockwise, starting top right:* Slytherin, Ravenclaw, Hufflepuff, Gryffindor), with the school's motto, which translates to "never tickle a sleeping dragon".[1]

Harry Potter and the Philosopher's Stone
Harry Potter and the Chamber of Secrets
Harry Potter and the Prisoner of Azkaban
Harry Potter and the Goblet of Fire
Harry Potter and the Order of the Phoenix

Figure 4: The Harry Potter article of 2011.

Conclusion: the Wayback Machine solves almost all of the problems mentioned in Section 1.2 by simply saving the complete output instead of trying to recreate it. It fulfils use case 1 described in Section 1.3. It's only drawback is its incompleteness, thus making it unsuitable for use case 2.

Websites:

- Internet Archive: http://www.archive.org
- Wayback Machine: http://www.archive.org/web/

2.2 Memento

Memento is a proposal for extending the HTTP protocol with a new header that adds the time we want to request, i.e. we cannot only get a resource but also a resource as it was at a previous point in time. A full explanation of the Memento system is given in the paper [4].

The Memento proposal has the drawback that, since it is an extension of the HTTP protocol, it requires both the extension of the server as well as the client. If it was there the user could simply browse not only from page to page on the Web, but also back in time at any given page.

For MediaWiki a server sided extension was developed that implements the HTTP extension. Furthermore the extension also takes care of the template resolution problem discussed in Section 1.2. It does not solve any of the other problems discussed in Section 1.1, thus the time travel does remain somehow incomplete (but is qualitatively better than the current solution, where each template is resolved against the current version instead of the version used back then). The extension is published under the GPL.

Consequently also a Firefox extension exists that implements the Memento protocol. On a page that is served using the Memento extension it allows the user to use a slider in order to view previous versions of the requested resource. The Memento extension for Firefox in combination with the server side Memento extension for MediaWiki allows to browse to previous versions of a MediaWiki page. The links at the end of this section point to a wiki with the Memento extension installed.

The possible usage of Memento on Wikipedia was rejected due to the expected increased performance costs. The MediaWiki database backend is optimized to offer current reversions. Accessing a big number of old template revisions in order to rebuild an old revision would be quite costly.

Conclusion: Memento would only solve one of the additional issues mentioned in Section 1.2, but for the prize of requiring all users who want to use it to install a plugin in their browser since the solution is based on the level of the HTTP protocol. It only aims at use case 1 described in Section 1.3 and can not be used for use case 2. Since Wikipedia does not nor is it expected to use the Memento extension, we cannot rely on this solution.

Websites:

- Memento: http://www.mementoweb.org/
- Memento Firefox extension http://addons.mozilla.org/en-US/firefox/addon/mementofox/
- Memento MediaWiki extension http://www.mediawiki.org/wiki/Extension:Memento
- Example wiki with Memento extension: http://wiki.2010.dev8d.org
- Discussion about the usage of Memento on Wikipedia: http://www.mail-archive.com/wikitech-l@lists.wikimedia.org/msg05707.html

2.3 JWPL

The Java WikiPedia Library (JWPL) is a Java-based library that enables the programmatic access to all information within Wikipedia. JWPL includes two subprojects that are relevant for the purpose of accessing time-contextual information: the Wikipedia Revision Toolkit and the JWPL TimeMachine.

JWPL allows accessing Wikipedia dumps, including the full site history. The naïve approach – i.e. just parse the full XML file and load it for further processing -- can have several disadvantages (if possible at all):

- Parsing the full XML revision is expensive and takes long. The English dump has Terabytes of data, and simply reading through it will take hours. Having random access to the history dump is extremely inefficient in any naïve way.
- A much more satisfying approach is to use the MediaWiki API, as it allows random access. But it requires internet access, as the Wikipedia MediaWiki API needs to be queried, increases disproportionally the server load, and has also a considerable time overhead.

For some use cases, one of these approaches can lead to a satisfying result. JWPL enables a non-naïve approach to the data locally, reducing the required disk space and still provide sufficiently quick random access mechanisms to all revisions.

The access to the data is improved by

- saving diffs between revisions instead of the complete revisions, thus decreasing required disk space to about 2% of the original requirement, and
- providing a simple Java API to enable simple programmatic access.

Besides the access to the full revision history, JWPL also provides a functionality to create a dump from any given point in time. That is, using JWPL we can specify a given point in time and, using a full dump, create a dump that was current at the given point in time. This allows creating well-defined datasets that can be used for creating reproducible results, a necessity for any good research. Thus research results become independent of the provisioning of any certain-dated current dump by the Wikimedia Foundation.

The JWPL TimeMachine even goes further by providing a facility to create a dump with snapshots following a specific interval between two timestamps, e.g. to get snapshots of every month for 2009 and 2010. The selection and usage of diffs reduces the amount of data, and, at the same time, speeds up the access to the data, and allows more efficient analysis of time-contextual data. The result can then be loaded into a MySQL database and be efficiently managed and used.

JWPL is provided open source under the GPL license and has an active and responsive mailing list. An extension by the project if necessary would thus be possible.

Conclusion: JWPL can be used for use case 2 from Section 1.3, but not for use case 1.

Websites:

- JWPL http://code.google.com/p/jwpl/

3 Time-contextual language corpora

In order to evaluate the relevance of time-contextual access to the data in Wikipedia, we have extended the creation of the Corpex natural language corpus created from Wikipedia with a temporal aspect.

3.1 Creating language corpora from Wikipedia

Here we quickly reiterate how we created the Corpex language corpora from Wikipedia. For full details and an extensive evaluation of the quality of the corpus, we refer to Deliverable 1.1.1 and the paper on Corpex [2].

We have taken the text of several Wikipedia language editions, cleansed it, and created corpora for 33 languages. In order to evaluate how viable these corpora are, we have calculated language models for the English Wikipedia, and compared it to widely used corpora. Since the English Wikipedia edition is far larger than any other — and size of a corpus is a crucial factor for its viability — we have also taken the Simple English Wikipedia edition, being smaller than many other language editions, and compared that as well. The results of this comparison support our assumption that the language models created from the corpora from other language editions have an acceptable quality.

We make the generated language models and the corpora available. The full data sets can be downloaded. The website also provides a novel, graphical corpus exploration tool – Corpex – not only over the newly created corpora that we report on here, but also usable for already established corpora like the Brown corpus.

Articles in Wikipedia are written using the MediaWiki syntax, a wiki syntax offering a flexible, but very messy mix of some HTML elements and some simple markup. There exists no proper formal definition for the MediaWiki syntax. It is hard to discern which parts of the source text of an article is actual content, and which parts provide further functions, like navigation, images, layout, etc. This introduces a lot of noise to the text.

We have filtered the article source code quite strictly, throwing away roughly a fourth of the whole content. This includes most notably all template calls, which are often used, e.g., to create infoboxes and navigational elements. The actual script that provides the filtering is available on the Corpex website as open source, so that it can be further refined. When exploring the corpus, one can easily see that quite some noise remains. We aim to further clean up the data and improve the corpora over time.

The content of the Wikipedia editions is provided as XML dumps. We have selected only the actual encyclopaedic articles, and not the numerous pages surrounding the project, including discussion pages for the articles, project management pages, user pages, etc., as we expect those to introduce quite some bias and idiosyncrasies. The table in Figure 1 contains the date when the XML dump was created, for reference and reproducibility of the results. Combined, we have processed around 75 gigabytes of data.

3.2 Adding time to Corpex

Out of the complete history dump of Wikipedia, we created snapshots (as described in Section 2.3). We created snapshots on January 1 of each year, and then used the same approach for filtering the text as for the normal Corpex. The results were lists of words and frequencies. The raw lists are made available on the Corpex website.

For the results we decided to try the full corpus of Wikipedia, and – unlike for the normal Corpex – not to filter for only the main namespace. The first look at the results shows that this was not a good idea. Many terms represent the idiosyncrasy and terminology of Wikipedia: words like “edit”, “user”, “revert”, etc. are grossly overrepresented, as well as the user names of active contributors. We will recreate the temporal corpora using only the main namespace and release these results as well.

For simple English, we have the following ground data for the current, non-filtered corpus.

Table 1: Some ground statistics for the time-stamped corpora of the Simple English Wikipedia.

Year	Terms	New terms	Term growth	Words	Absolute word growth	Relative growth
2003	33,693	33,693	-	291,615	291,615	-
2004	42,694	9,041	+26.8%	452,292	160,677	+55.1%
2005	58,323	16,416	+38.5%	801,751	349,459	+77.3%
2006	81,570	21,137	+41.4%	1,540,020	738,269	+92.1%
2007	119,009	40,536	+49.7%	3,719,894	2,179,874	+141.5%
2008	174,094	60,426	+50.8%	6,797,752	3,077,858	+82.7%
2009	244,014	75,987	+43.6%	11,341,742	4,543,990	+66.8%

Terms gives the number of unique words in the given year, *new terms* is the number of unique words that have not been present in the previous year. Note that this is not necessarily the difference between terms in one year to the next as terms can also be removed (e.g. from 2008 to 2009, the term 'megastrike', which appeared 11 times in the 2008 corpus, was purged from the corpus). The *term growth* is the relative growth in the given year in terms. *Words* gives the number of non-unique occurrences of all terms, the *absolute word growth* is the difference to the previous year, and the *relative growth* gives the percentage compared to the previous year in words.

3.3 First results of the analysis

As discussed, due to the size of the edit history, the analysis is running longer than expected. We will continue to run the analysis, and will publish the results afterwards. For now, we have available the processed data from the Simple English Wikipedia from 2003-2009, i.e. from its beginning up to two years ago, given non-filtered results. A few overview numbers are given in the previous section.

Table 2: Example words and their distribution over time.

Term	2003	2004	2005	2006	2007	2008	2009
the	1,806	9,536	28,223	69,191	174,000	343,784	583,793
	6,190	21,100	35,200	44,900	46,800	50,600	51,500
		48,100	53,500	55,500	48,000	55,200	52,800
edit	49	149	200	472	2,845	4,739	7,554
	168	329	249	306	765	697	666
		622	146	368	1,090	615	619
revert	2	8	15	25	232	501	902
	6.86	17.7	18.7	16.2	62.4	73.7	79.5
		37.4	20.0	13.5	95.0	87.4	88.2
iphone	2	2	2	2	2	16	32
	6.86	4.42	2.49	1.30	0.538	2.35	2.82
		0	0	0	0	4.55	3.52

africa	20	48	147	334	592	968	1.493
	68.6	106	183	216	159	142	131
		174	273	253	118	122	116
europe	13	75	216	404	918	1.684	2.475
	44.6	166	269	262	247	248	218
		386	403	255	236	249	174
merkel	0	0	0	6	16	24	37
	0	0	0	3.90	4.30	3.53	3.26
		0	0	8.13	4.59	2.60	2.86
diversity	0	0	3	4	11	39	69
	0	0	3.74	2.60	2.96	5.74	6.08
		0	8.58	1.35	3.21	9.10	6.60

The first number gives the absolute count, the second number gives the count per million words, and the third number gives the count per million *within the new words only that have been added in the previous year*. A first look at the numbers shows that there are stable words and words that are more volatile and that unveil a change in interest over time. We are working towards classifying these different terms accordingly, and to understand what these differences mean.

We are preparing an extension of the Corpex website that will allow the easy access to the Corpex time corpus. Words that are typed will show their usage over time similar to the numbers given above.

4 Conclusions

This deliverable has demonstrated the necessity for a time-contextual access to wiki-like items. We have described two use cases:

1) for users to view previous versions of a page, and
2) for analysing the content of a wiki at a specific point of time.

We have seen that the current solutions provided by Wikipedia are insufficient for both use cases.

We have surveyed existing solutions, and discussed their benefits and their issues. The three solutions are:

1) the Internet Archive's Wayback Machine, which serves the first use case well, especially in combination with the normal history feature of Wikipedia,
2) Memento, an extension to HTTP, which is prototypically implemented for MediaWiki but which has some major drawbacks that make it not suitable for our use cases, and
3) JWPL, a Java library that allows us to serve use case 2.

We have decided that we do not need to create a new solution for the given problem, since the combination of the normal Wikipedia history view, the Wayback Machine, and JWPL with the Wikipedia dumps is sufficient for our use cases.

The effort that we spared we invested in creating an infrastructure to extract a time-contextual language corpus from Wikipedia. The extraction is still ongoing, but first results are presented in this deliverable. We will publish a follow-up about the final results, and also extend the Corpex website with the complete dataset as soon as it is available.

References

[1] http://www.render-project.eu

[2] *Corpex:* Vrandecic, D., Sorg, P., Studer, R. (2011), Language Resources extracted from Wikipedia, in Proceedings of the international conference on Knowledge Capute (K-CAP)

[3] *JWPL:* Ferschke, O., Zesch, T. und Gurevych, I. (2011), Wikipedia Revision Toolkit: Efficiently Accessing Wikipedia's Edit History, in Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. System Demonstrations

[4] *Memento:* van de Sompel, H., Nelson, M., Sanderson, R., Balakireva, L., Ainsworth, S., Shankar, H. (2009), Memento: Time Travel for the Web, in arXiv:0911.1112v2

