What you say is who you are. How open government data facilitates profiling politicians

Maarten Marx and Arjan Nusselder

ISLA, Informatics Institute, University of Amsterdam
Science Park 107 1098XG Amsterdam, The Netherlands

Abstract. A system is proposed and implemented that creates a language model for each member of the Dutch parliament, based on the official transcripts of the meetings of the Dutch Parliament. Using expert finding techniques, the system allows users to retrieve a ranked list of politicians, based on queries like news messages.
The high quality of the system is due to extensive data cleaning and transformation which could have been avoided when it had been available in an open machine readable format.

1 Introduction

The Internet is changing from a web of documents into a web of objects. Open and interoperable (linkable) data are crucial for web applications which are build around objects. Examples of objects featuring prominently is (mashup) websites are traditional named entities like persons, products, organizations [6,4], but also events and unique items like e.g. houses.

The success of several mashup sites is simply due to the fact that they provide a different grouping of already (freely) available data. Originally the data could only be grouped by documents; the mashup allows for groupings by objects which are of interest in their specific domain.

Here is an example from the political domain. Suppose one wants to know more about *Herman van Rompuy*, the new EU "president" from Belgium. Being a former member of the Belgium parliament and several governments, an important primary source of information are the parliamentary proceedings. Most states serve these proceedings on the web with some search functionality. But in most states the information is organized in a document-centric way. One can retrieve documents containing the verbatim notes of a complete parliamentary day. It is simply not possible to state the query

> return all speeches made by person X on topic Y.

Of course this information is available in the proceedings but most often not in machine readable form. Austria, the EU and the British website `theyworkforyou.com` show that it is possible and useful to provide entrance to parliamentary information starting from politicians.

Contribution of this paper We build a people search engine using an out of the box search application (Indri `http://www.lemurproject.org/indri`). We did an extensive evaluation using standard information retrieval metrics which showed that its performance compares to the state of the art. We achieved this performance because of an elaborate and very careful data extraction and data cleaning which would not be needed if the data had been available in an open format (as provided by e.g. the EU parliament).

The purpose of this paper is thus to give a concrete example of the power and impact of open, in our case governmental, data. The paper is organized as follows. We first briefly introduce the field of people search. Then we describe the politician search engine that was created. We describe the data and the used retrieval method and end with an extensive evaluation.

2 People search

The field within information retrieval which is concerned with information needs for concrete persons is called *People Search* [1]. The most basic application consists of a Google style search box in which a user can paste text after which a list of references to persons is returned. The list is ranked by relevance of the persons to the topic expressed by the input text. An important application is expertise retrieval: Within an organization, search for experts which can help on a user-provided topic. The matching functionality of dating sites can also be seen as a form of people search.

We now describe in a simplified form the technique behind most expertise retrieval applications participating in the TReC expertise retrieval task []. The first step is data collection. Typically a crawl of some companies' intranet is available, sometimes with personal data as emails etc. Using named entity extraction techniques, occurrences of persons in the text are recognized and reconcilliated to items in a given list of persons. Data deduplication [2] is the main bottleneck in this step.

Then the system creates a model of each person based on the textual and structural content of the crawl. Simply said, each person is represented by one long text document consisting of pieces of text strongly related to that person. The problem here is to determine which parts of the texts in the crawl should be connected to which person.

It should be clear that these problems can be simply avoided for parliamentary data if 1) a unique naming convention is used, and 2) what is being said by whom is structurally made clear. An example of an XML data format in which this is the case is given in Figure 1. Note that the text is a small speech given by the chairman, who is called Jerzy Buzek. Using the MPid attribute he is uniquely linked to a concrete person.[1]

[1] The person can be found by following this link `http://www.europarl.europa.eu/members/public/geoSearch/view.do?language=EN&partNumber=1&country=PL&id=28269`

```
<speech docno="3-010"
        MPid="28269"
        MPname="Przewodniczcy.">
 <p docno="3-010-1">
   <stage-direction>Przewodniczcy. </stage-direction>
   Kolejnym punktem porzdku dziennego s owiadczenia
   Rady i Komisji dotyczce przygotowania posiedzenia
   Rady Europejskiej w dniach 10 i 11 grudnia 2009 r.
 </p>
</speech>
```

Fig. 1. Piece of well structured parliamentary proceeding.

3 Matching politicians to news articles

We build a retrieval system which performs the following task:

> given a news article, find those members of parliament which have a strong interest in the topic of the article. Rank them by the strength of that interest.

To match politicians to news we used the people search approach described above. We used the parliamentary proceedings to build a profile of each politician. A description of the data is given in section 4. The resulting system can be seen as answering the question: “Given the words spoken in parliament by a politician, how well does she match a given text?”

Our approach to the retrieval of politicians is based largely on work done by Balog [1]. We used his *Model 1*, which describes the idea of representing experts –politicians in our case– as single documents.This model itself is based on language modelling techniques [7][5].

4 Data

We created a language model of each sitting member of the Dutch parliament. As textual input data we took the parliamentary proceedings which record everything being said in parliament. Through the PoliticalMashup project [3], this data is available in XML in a format which is excellent for our task: every word is annotated with the name of its speaker, her party and the date.

Besides these primary data sources we used biographical data about our politicians available at `www.parlement.com`.

5 Method

What needs to be expressed somehow, is the chance that a politician is knowledgeable on –or at least interested in– the topic expressed by a query. To do so, each politician must be represented with a profile. We first define such a profile as a document in which all text related to that politician is concatenated. This way, the politician–topic matching problem can be reduced to an instance of ranked document retrieval. To calculate the probabilities and ranking, the query is compared to all politicians, each represented as a language model of the concatenation of the related texts.

The measure used for comparison is the Kullback-Leibler divergence. We take $Q : Word \rightarrow Wordcount$ as the function over the words in the query, and $P : Word \rightarrow Wordcount$ as the function over the words in a document representing a politician. The basic formula to calculate the chance of a query given a politician is expressed in equation (1).

$$KL(Q|P) = \sum_i Q(i) \log \frac{Q(i)}{P(i)} \tag{1}$$

The result of a query is a ranked list of document identifiers, corresponding to the politician the texts belong to. To create an accessible and usable interface, the results are embedded in a block of additional information. At the time of writing, an interface is available at `http://zookma.science.uva.nl/politiciansearch/search.php`

For the actual implementation, the Lemur Toolkit was used.[2] The important `Lemur` parameters are *Simple KL* as retrieval model, and for smoothing a *Dirichlet prior* set at the total number of word types.

Some additional ideas focusing more on the presentation of the results have been implemented. It is possible to not only collect texts on a per person basis, but also split the aggregations on a temporal or party level. Using a log-likelihood comparison, politicians can then be described as opposed to other politicians, or in a specific time-frame. Extensions like these could improve the usefulness of a system, but are left for future evaluation.

6 Evaluation

To see how well our approach performs, an experimental evaluation similar to the TREC 2005 W3C enterprise search task was devised.[3] The Dutch parliament has 23 committees, each focused on a policy topic, roughly corresponding to the existing ministeries[4]. Each committee consists of about eight to twenty-five members, and an equal or smaller number of reserve members. For each committee its name, a short description and its members (all MP's) are known. We used the both the committee names and their descriptions as topics. A result (i.e., a politician) is correct ("relevant") on a topic iff it is an active member of the committee described by that topic (reserve members were not counted). The

[2] See: `http://www.lemurproject.org`

[3] See: `http://trec.nist.gov/`

[4] See: `http://www.tweedekamer.nl/kamerleden/commissies/index.jsp`

total number of candidates is 150, which is the number of current members of parliament.

Thus we do two evaluation runs, one with the names of the committees as topics, and one with the descriptions of the committees. Committee names consist of 1 to 5 words (excluding stopwords); descriptions are between 500 and 1000 words. For instance, the description for the finance committee is 638 words (including stopwords).[5] Table 1 gives two examples of committee names; Table 2 contains a part of the description of committee with topic id 6.

These longer descriptions match the purpose of our recommendation system more closely.

6	Commissie voor de Verzoekschriften en de Burgerinitiatieven
8	Financien

Table 1. Names of topics 6 and 8, as they were used as query-text for the evaluation.

Comissie voor de Verzoekschriften en de Burgerinitiatieven Commissie Verzoekschriften en Burgerinitiatieven De commissie voor de Verzoekschriften en de Burgerinitiatieven heeft twee taken: het voorbereiden van een beslissing van de Kamer over een individuele aangelegenheid (waar een burger in een verzoekschrift om heeft gevraagd) en het voorbereiden van een beslissing van de Kamer over de ontvankelijkheid van een burgerinitiatief. . .

Table 2. Beginning of the description of topic 6.

Results. We measured the mean average precision (MAP) and precision at 10 (P@10) over two times 23 topics. The results are in Table 3.

Precision at ten is taken as an appropriate measure for two reasons. First, some committees have little more than ten members, which would make precision over ten difficult to evaluate. Second, the intended use of the application foresees a human-readable result set. Figure 2 shows the P@10 for each topic for both evaluation runs (full description and the committee-name only), with the topics ordered by their P@10 for the description run. Figure 3 additionally shows the MAP score of each topic, ordered by topic id, for the full descriptions topics.

For the majority of topics –or committees– more than 6 from the first ten results were correct when we used the full description. Looking at figure 2, some possible problems can be identified. Query 8 shows a large discrepancy between the full description and the name only. This may be due to the fact that the topic –just the singe word finance– can be and probably is used in virtually all contexts. The full text of the finance topic is descriptive enough to allow for a match between politicians focused on this area and the committee. The fact

[5] The description can be found at `http://www.tweedekamer.nl/kamerleden/commissies/FIN/sub/index.jsp`.

	MAP	**P@10**
committee names	.38	.48
committee descriptions	.44	.56

Table 3. MAP and P@10 of our experiments.

that almost all politicians will talk about financial issues however, could make the committee name by itself insufficient. Because the focus of the application lies on a search for more verbose text, this is not necessarily a problem.

Query 6 performs worse both with the full description and only the committee name. Several problems may be the cause of this. First, the committee itself consists –as an exception– of only eight members, which makes it harder to correctly retrieve the correct politicians. Also the topic of the committee is relatively new as compared to others, meaning there is probably less data available to create a profile that acknowledges this specific interest of the members. Third, the topic is pretty vague and seems rather specialized.

A small evaluation (3 topics) which mimics exactly the use-case in mind (finding politicians likely to be interested in a news-story) gave even better results: all topics got a P@10 of .6 or higher. These results can be found at `http://zookma.science.uva.nl/politiciansearch/search.php`.
Here the reader can also evaluate the system herself. Interesting queries are "ik" (*I*), "Nederland" (*The Netherlands*) and "vrede" *peace.*

7 Conclusion

The high precision scores obtained by a baseline implementation show the importance of well presented data. Making data available in an open and linkable format using unique identifiers makes it much easier to build robust systems.

References

1. K. Balog. *People Search in the Enterprise.* PhD thesis, University van Amsterdam, September 2008.
2. X. Dong, A. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In *Proc. SIGMOD*, pages 85–96, 2005.
3. T. Gielissen and M. Marx. Exemelification of parliamentary debates. In *Proc. DIR*, 2009.
4. M. Hearst. *Search User Interfaces.* Cambridge University Press, 2009.
5. D. Hiemstra. *Using Language Models for Information Retrieval.* PhD thesis, University of Twente, 2001.
6. C. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval.* Cambridge University Press, 2008.
7. J. Ponte and W. Croft. A language modelling approach to information retrieval. *Proc. SIGIR '98*, 1998.

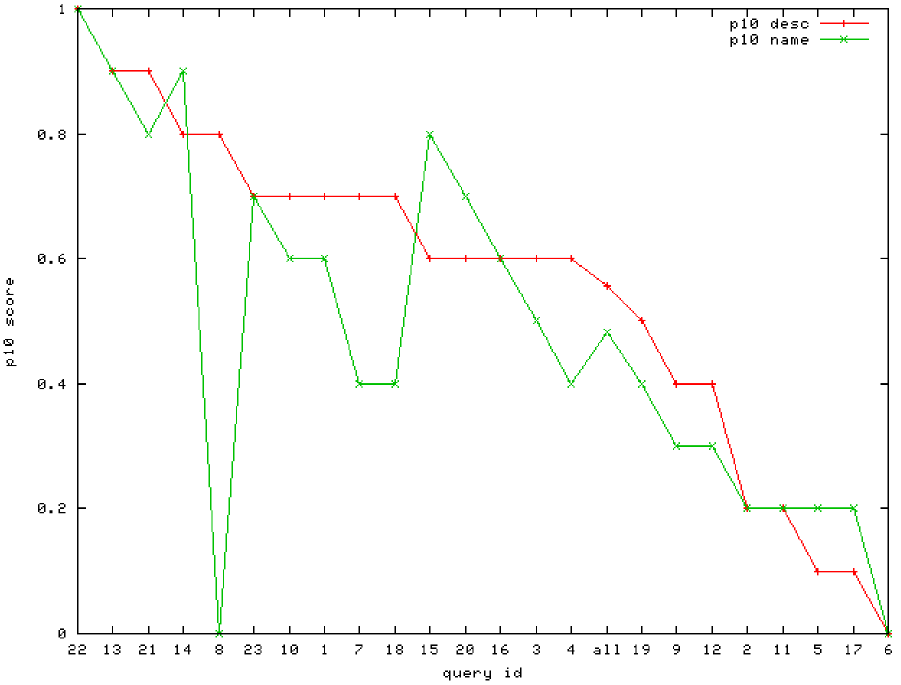


Fig. 2. Precision at ten for the full description (desc) and the committee-names (name).

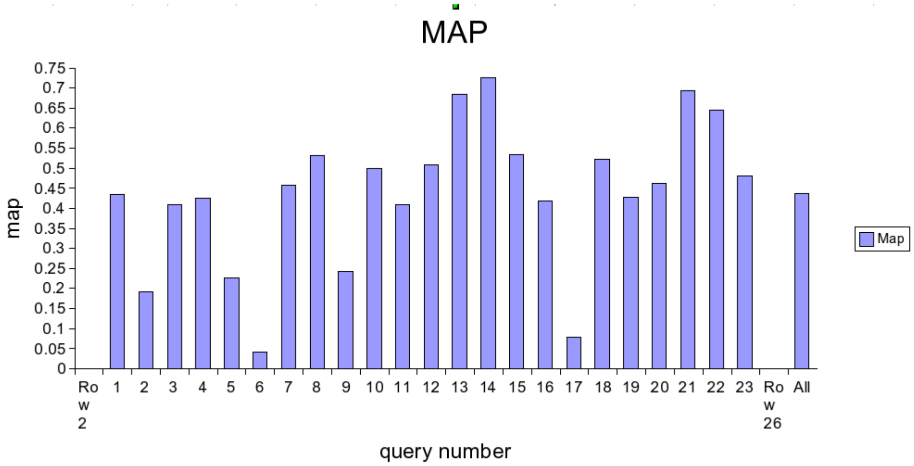


Fig. 3. Mean average precision for each full text query.

Collaborative Structuring of Knowledge by Experts and the Public

Tom Morris[1] and Daniel Mietchen[2]

[1] http://www.citizendium.org/User:Tom_Morris
[2] http://www.citizendium.org/User:Daniel_Mietchen
Correspondence: Daniel.Mietchen (at) uni-jena (dot) de

Abstract. There is much debate on how public participation and expertise can be brought together in collaborative knowledge environments. One of the experiments addressing the issue directly is Citizendium. In seeking to harvest the strengths (and avoiding the major pitfalls) of both user-generated wiki projects and traditional expert-approved reference works, it is a wiki to which anybody can contribute using their real names, while those with specific expertise are given a special role in assessing the quality of content. Upon fulfillment of a set of criteria like factual and linguistic accuracy, lack of bias, and readability by non-specialists, these entries are forked into two versions: a stable (and thus citable) approved ”cluster” (an article with subpages providing supplementary information) and a draft version, the latter to allow for further development and updates. We provide an overview of how Citizendium is structured and what it offers to the open knowledge communities, particularly to those engaged in education and research. Special attention will be paid to the structures and processes put in place to provide for transparent governance, to encourage collaboration, to resolve disputes in a civil manner and by taking into account expert opinions, and to facilitate navigation of the site and contextualization of its contents.

Key words: open knowledge, open education, open science, open governance, wikis, expertise, Citizendium, Semantic Web

1 Introduction

> *Science is already a wiki if you look at it a certain way. It's just a highly inefficient one – the incremental edits are made in papers instead of wiki-space, and significant effort is expended to recapitulate existing knowledge in a paper in order to support the one to three new assertions made in any one paper.*
>
> John Wilbanks [21]

There are many ways to structure knowledge, including collaborative arrangements of digital documents. Only a limited number of the latter ones have so far been employed on a larger scale. Amongst them are wikis – online platforms which allow the aggregation, interlinking and updation of diverse sets of knowledge in an Open Access manner, i.e. with no costs to the reader.

1.1 Wikis as an example of public knowledge environments online

As implied by the introductory quote, it is probably fair to say that turning science (or any system of knowledge production, for that matter) into a wiki (or a set of interlinked collaborative platforms) would make research, teaching and outreach much more transparent, less prone to hype, and more efficient. Just imagine you had a time slider and could watch the history of research on general relativity, plate tectonics, self-replication, or cell division unfold from the earliest ideas of their earliest proponents (and opponents) onwards up to you, your colleagues, and those with whom you compete for grants. So why don't we do it?

Traditionally, given the scope of a particular journal, knowledge about specialist terms (which may describe completely non-congruent concepts in different fields), methodologies, notations, mainstream opinions, trends, or major controversies could reasonably be expected to be widespread amongst the audience, which reduced the need to redundantly say and then repeat the same things all over again and again (in cross-disciplinary environments, there is a higher demand for proper disambiguation of the various meanings of a term). Nonetheless, redundancy is still quite visible in journal articles, especially in the introduction, methods, and discussion sections and the abstracts, often in a way characteristic of the authors (such that services like eTBLAST and JANE can make qualified guesses on authors of a particular piece of text, with good results if some of the authors have a lot of papers in the respective database, mainly PubMed, and if they have not changed their individual research scope too often in between).

A manuscript well-adapted to the scope of one particular journal is often not very intelligible to someone outside its intended audience, which hampers cross-fertilization with other research fields (we will get back to this below). When using paper as the sole medium of communication there is not much to be done about this limitation. Indeed, we have become so used to it that some do not perceive it as a limitation at all. Similar thoughts apply to manuscript formatting. However, the times when paper alone reigned over scholarly communication have certainly passed, and wiki-like platforms provide for simple and efficient means of storing information, updating it and embedding it into a wider context.

Cross-field fertilization, for example, is crucial with respect to interdisciplinary research projects, digital libraries and multi-journal (or indeed cross-disciplinary) bibliographic search engines (e.g. Google Scholar), since these dramatically increase the likelihood of, say, a biologist stumbling upon a not primarily biological source relevant to her research (think shape quantification or growth curves, for instance). What options do we have to systematically integrate such cross-disciplinary hidden treasures with the traditional intra-disciplinary background knowledge and with new insights resulting from research?

The by now classical example of a wiki environment are the Wikipedias, a set of interlinked wikis in multiple languages where basically anyone can edit any page, regardless of subject matter expertise or command of the respective language. As a consequence of this openness, the larger Wikipedias have a serious problem with vandalism: take an article of your choice and look at its history page for reverts - most of them will be about neutralizing subtle or blunt forms of destructive edits that do nothing to improve the quality of the articles, but may reduce it considerably. Few of these malicious edits persist for long [14], but finding and fixing them takes time that could better be spent on improving articles. This is less of an issue with more popular topics for which large numbers of volunteers may be available to correct ”spammy” entries but it is probably fair to assume that most researchers value their time too much to spend it on repeatedly correcting information that had already been correctly entered. Other problems with covering scientific topics at the Wikipedias include the nebulous notability criteria which have to be fulfilled to avoid an article being deleted, and the rejection of ”original research” in the sense of not having been peer reviewed before publication. Despite these problems, one scientific journal – RNA Biology – already requires an introductory Wikipedia article for a subset of papers it is to publish [16].

Peer review is indeed a central aspect of scholarly communication, as it paves the way towards the reproducibility that forms one of the foundations of modern science. Yet we know of no compelling reason to believe that it works better before than after the content concerned has been made public (doing it beforehand was just a practical decision in times when journal space was measured in paper pages), while emerging movements like Open Notebook Science – where claims are linked directly to the underlying data that are being made public as they arise – represent an experiment in this direction whose initial results look promising and call into question Wikipedia’s ”no original research” as a valid principle for generating encyclopaedic content.

Although quite prominent at the moment, the Wikipedias are not the only wikis around, and amongst the more scholarly inclined alternatives, there are even a number of wiki-based journals, though usually with a very narrow scope and/or a low number of articles. On the other hand, Scholarpedia (which has classical peer review and an ISSN and may thus be counted as a wiki journal, too [17]), OpenWetWare [12], Citizendium [2] and the Wikiversities [20] are cross-disciplinary and structured (and of a size, for the moment) such that vandalism and notability are not really a problem. With minor exceptions, real names are required at the first three, and anybody can contribute to entries about anything, particularly in their fields of expertise. None of these is even close to providing the vast amount of context existing in the English Wikipedia but the difference is much less dramatic if the latter were broken down to scholarly useful content. Out of these four wikis, only OpenWetWare is explicitly designed to harbour original research, while the others allow different amounts thereof. Furthermore, a growing number of yet more specialized scholarly wikis exist (e.g. WikiGenes [18], the Encyclopedia of Earth [8], the Encyclopedia of Cosmos [7], the Dispersive PDE Wiki [6], or the Polymath Wiki [13]), which can teach us about the usefulness of wikis within specific academic fields.

2 The Citizendium model of wiki-based collaboration

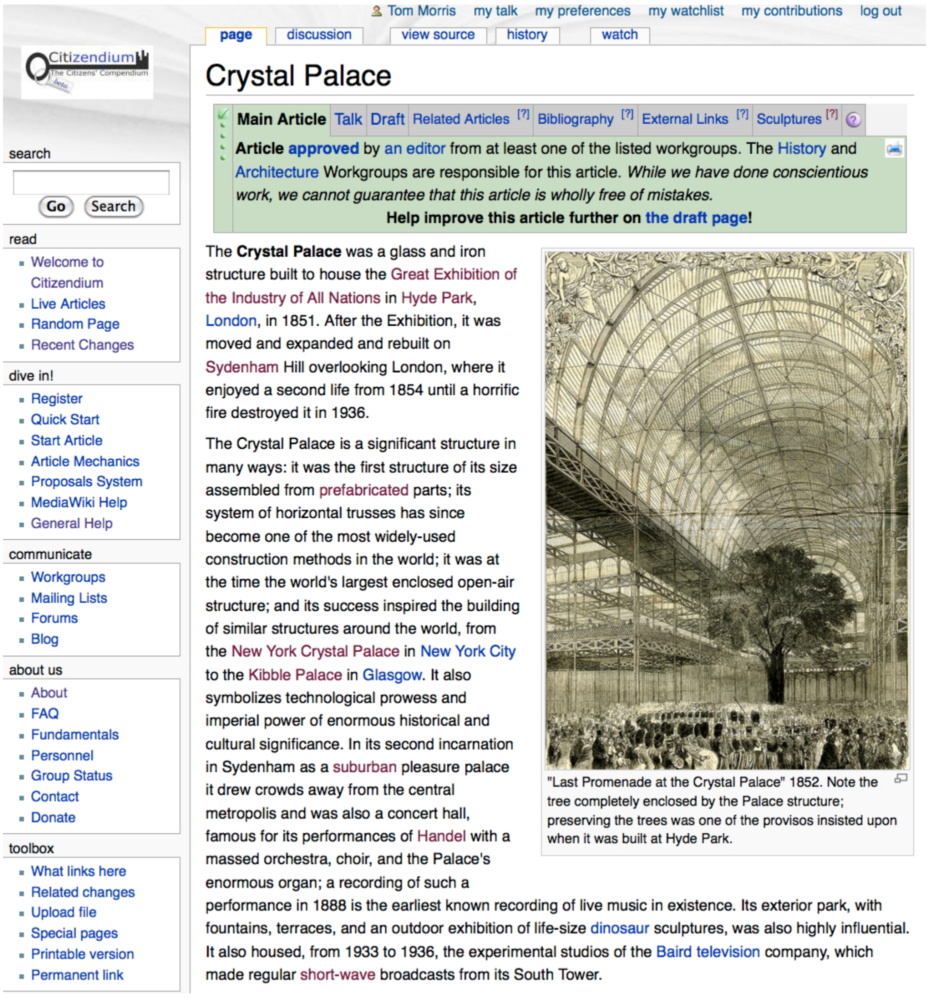
Despite the above-mentioned tensions between public participation and expertise in the collaborative structuring of knowledge, it is not unreasonable to expect that these can be overcome by suitably designed public knowledge environments, much like Citizen Science projects involve the public in the generation of scientific data. One approach at such a design is represented by Citizendium. The founder of Citizendium – Larry Sanger – is the co-founder of Wikipedia. The two projects share the common goal of providing free knowledge to the public, they are based on variants of the same software platform, and they use the same Creative Commons-Attribution-Share Alike license [4]. Yet they differ in a number of important ways, such that Citizendium can be seen as composed of a Wikipedia core (stripped down in terms of content, templates, categories and policies), with elements added that are characteristic of the other wiki environments introduced above: A review process leading to stable versions (as at Scholarpedia), an open education environment (as at Wikiversity) and an open research environment (as at OpenWetWare). Nonetheless, assuming that the reader is less familiar with these three latter environments, we will follow previous commenters and frame the discussion of Citizendium in terms of properties differentiating it from Wikipedia, and specifically the latter's English language branch [19].

2.1 Real names

The first of these is simply an insistence on real names. While unusual from a Wikipedia perspective, this is custom in professional environments, including traditional academic publishing and some of the above-mentioned wikis, e.g. Scholarpedia and Encyclopedia of Earth. It certainly excludes a number of legitimate contributors who prefer to remain anonymous but otherwise gives participants accountability and allows to bring in external reputation to the project.

2.2 Expert guidance

To compose and develop articles and to embed them in the multimedial context of a digital knowledge environment, expert guidance is important. Of course, many experts contribute to Wikipedia, and the Wikipedias in turn have long started to actively seek out expert involvement, yet the possibility to see their edits overturned by anonymous users that may lack even the most basic education in that field keeps professionals away from spending their precious time on such a project. The Citizendium approach of verifying expertise takes a different approach – sometimes termed "credentialism" – that rests on a common sense belief that some people do know more than others: it is sometimes the case that the thirteen-year-old kid in Nebraska does know more than the physics professor. But most of the time, at least when matters of physics are concerned, this is not the case. The role the experts have at Citizendium is not, as frequently



Crystal Palace

Article approved by an editor from at least one of the listed workgroups. The History and Architecture Workgroups are responsible for this article. *While we have done conscientious work, we cannot guarantee that this article is wholly free of mistakes.*

Help improve this article further on the draft page!

The **Crystal Palace** was a glass and iron structure built to house the Great Exhibition of the Industry of All Nations in Hyde Park, London, in 1851. After the Exhibition, it was moved and expanded and rebuilt on Sydenham Hill overlooking London, where it enjoyed a second life from 1854 until a horrific fire destroyed it in 1936.

The Crystal Palace is a significant structure in many ways: it was the first structure of its size assembled from prefabricated parts; its system of horizontal trusses has since become one of the most widely-used construction methods in the world; it was at the time the world's largest enclosed open-air structure; and its success inspired the building of similar structures around the world, from the New York Crystal Palace in New York City to the Kibble Palace in Glasgow. It also symbolizes technological prowess and imperial power of enormous historical and cultural significance. In its second incarnation in Sydenham as a suburban pleasure palace it drew crowds away from the central metropolis and was also a concert hall, famous for its performances of Handel with a massed orchestra, choir, and the Palace's enormous organ; a recording of such a performance in 1888 is the earliest known recording of live music in existence. Its exterior park, with fountains, terraces, and an outdoor exhibition of life-size dinosaur sculptures, was also highly influential. It also housed, from 1933 to 1936, the experimental studios of the Baird television company, which made regular short-wave broadcasts from its South Tower.

"Last Promenade at the Crystal Palace" 1852. Note the tree completely enclosed by the Palace structure; preserving the trees was one of the provisos insisted upon when it was built at Hyde Park.

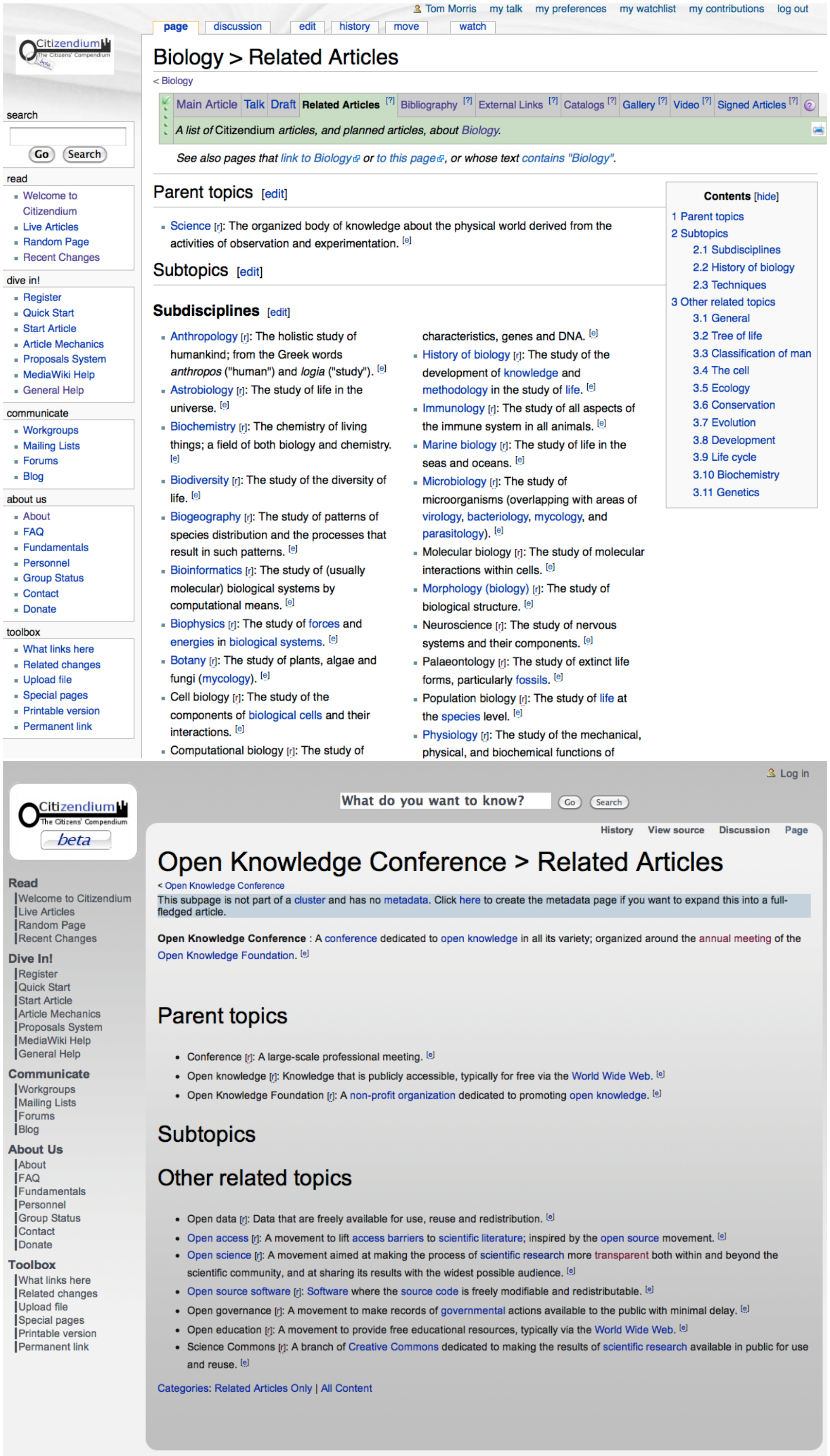
Fig. 1. Screenshot of the main page of the [[Crystal Palace]] cluster while logged in using a monobook skin that is the default at Wikipedia. It shows the *green cluster bar* that indicates that the page has been approved and links to all the supages. Also visible is the *status indicator* (green dots on the left topped by green tick), mention of *"an editor"* to distinguish the number of editors involved (some pages can be approved by one rather than three editors), links to the workgroups which have approval rights for the article (in this case: the *History and Architecture Workgroups*), a prominent *disclaimer* (unapproved articles have a much strong disclaimer), and links to the 'unstable' *draft* version of the article which any registered contributor can update. Like traditional encyclopaedic environments, Citizendium does not require every statement to be referenced, in the interest of narrative flow.

stated in external comments, that of a supreme leader who is allowed to exercise his will on the populace. On the contrary, it is much more about guiding. We use the analogy of a village elder wandering around the busy marketplace [15] who can resolve disputes and whom people respect for their mature judgement, expertise and sage advice. Wikipedia rejects "credentialism" in much the same way that the Internet Engineering Task Force (IETF) does. David Clark summarised the IETF process thusly [3]: "We reject kings, presidents and voting. We believe in rough consensus and running code." In an open source project, or an IETF standardisation project, one can decide a great many of the disputes with reference to technical reality: the compiler, the existing network protocols etc. If the code doesn't compile, think again. For rough consensus to happen under such circumstances, one needs to get the people together who have some clear aim in mind: getting two different servers to communicate with one another. The rough consensus required for producing an encyclopaedia article is different – it should attempt to put forward what is known, and people disagree on this to a higher degree than computers do on whether a proper connection has been established. It is difficult to get "rough consensus, running code" when two parties are working on completely different epistemological standards. At this point, one needs the advice of the village elderly who will vet existing content and provide feedback on how it can be expanded or otherwise improved. Upon fulfillment of a set of criteria like factual and linguistic accuracy, lack of bias, and readability by non-specialists, these vetted entries are forked into two versions: a stable (and thus citable) approved "cluster" (an article with subpages providing supplementary information) and a draft version, the latter to allow for further development and updates (cf. Fig. 1).

The respect for experts because of their knowledge of facts is only part of the reasoning: the experts point out and correct factual mistakes, but they also help to guide the structuring of content within an article and by means of the subpages. The experts bring with them the experience and knowledge of years of in-depth involvement with their subject matter, and the project is designed to make best use of this precious resource, while still allowing everyone to participate in the process. Of course, experts are likewise free to bring in content, be it within their specialty or in other areas, where others take over the guiding role. The Citizendium can also host 'Signed Articles', which are placed in a subpage alongside the main article. A Signed Article is an article on the topic described by a recognised expert in the field, but can express opinions and biases in a way that the main article ought not to.

2.3 Contextualization

Citizendium attempts to structure knowledge in a different way. Each article on Citizendium can make comprehensive use of Subpages, i.e. pages providing additional information that are subordinate to an article's page. Some of these – e.g. the Sculptures subpage in Fig. 1 – are similar to but more flexible than the supplementary online materials now being published routinely along scholarly articles. Two subpages types are different, with keywords and running title being



Biology > Related Articles

< Biology

A list of Citizendium *articles, and planned articles, about Biology.*

See also pages that link to Biology or to this page, or whose text contains "Biology".

Parent topics [edit]

- Science [r]: The organized body of knowledge about the physical world derived from the activities of observation and experimentation. [e]

Subtopics [edit]

Subdisciplines [edit]

- Anthropology [r]: The holistic study of humankind; from the Greek words *anthropos* ("human") and *logia* ("study"). [e]
- Astrobiology [r]: The study of life in the universe. [e]
- Biochemistry [r]: The chemistry of living things; a field of both biology and chemistry. [e]
- Biodiversity [r]: The study of the diversity of life. [e]
- Biogeography [r]: The study of patterns of species distribution and the processes that result in such patterns. [e]
- Bioinformatics [r]: The study of (usually molecular) biological systems by computational means. [e]
- Biophysics [r]: The study of forces and energies in biological systems. [e]
- Botany [r]: The study of plants, algae and fungi (mycology). [e]
- Cell biology [r]: The study of the components of biological cells and their interactions. [e]
- Computational biology [r]: The study of characteristics, genes and DNA. [e]
- History of biology [r]: The study of the development of knowledge and methodology in the study of life. [e]
- Immunology [r]: The study of all aspects of the immune system in all animals. [e]
- Marine biology [r]: The study of life in the seas and oceans. [e]
- Microbiology [r]: The study of microorganisms (overlapping with areas of virology, bacteriology, mycology, and parasitology). [e]
- Molecular biology [r]: The study of molecular interactions within cells. [e]
- Morphology (biology) [r]: The study of biological structure. [e]
- Neuroscience [r]: The study of nervous systems and their components. [e]
- Palaeontology [r]: The study of extinct life forms, particularly fossils. [e]
- Population biology [r]: The study of life at the species level. [e]
- Physiology [r]: The study of the mechanical, physical, and biochemical functions of

Contents [hide]

1 Parent topics
2 Subtopics
 2.1 Subdisciplines
 2.2 History of biology
 2.3 Techniques
3 Other related topics
 3.1 General
 3.2 Tree of life
 3.3 Classification of man
 3.4 The cell
 3.5 Ecology
 3.6 Conservation
 3.7 Evolution
 3.8 Development
 3.9 Life cycle
 3.10 Biochemistry
 3.11 Genetics

Open Knowledge Conference > Related Articles

< Open Knowledge Conference

This subpage is not part of a cluster and has no metadata. Click here to create the metadata page if you want to expand this into a full-fledged article.

Open Knowledge Conference : A conference dedicated to open knowledge in all its variety; organized around the annual meeting of the Open Knowledge Foundation. [e]

Parent topics

- Conference [r]: A large-scale professional meeting. [e]
- Open knowledge [r]: Knowledge that is publicly accessible, typically for free via the World Wide Web. [e]
- Open Knowledge Foundation [r]: A non-profit organization dedicated to promoting open knowledge. [e]

Subtopics

Other related topics

- Open data [r]: Data that are freely available for use, reuse and redistribution. [e]
- Open access [r]: A movement to lift access barriers to scientific literature; inspired by the open source movement. [e]
- Open science [r]: A movement aimed at making the process of scientific research more transparent both within and beyond the scientific community, and at sharing its results with the widest possible audience. [e]
- Open source software [r]: Software where the source code is freely modifiable and redistributable. [e]
- Open governance [r]: A movement to make records of governmental actions available to the public with minimal delay. [e]
- Open education [r]: A movement to provide free educational resources, typically via the World Wide Web. [e]
- Science Commons [r]: A branch of Creative Commons dedicated to making the results of scientific research available in public for use and reuse. [e]

Categories: Related Articles Only | All Content

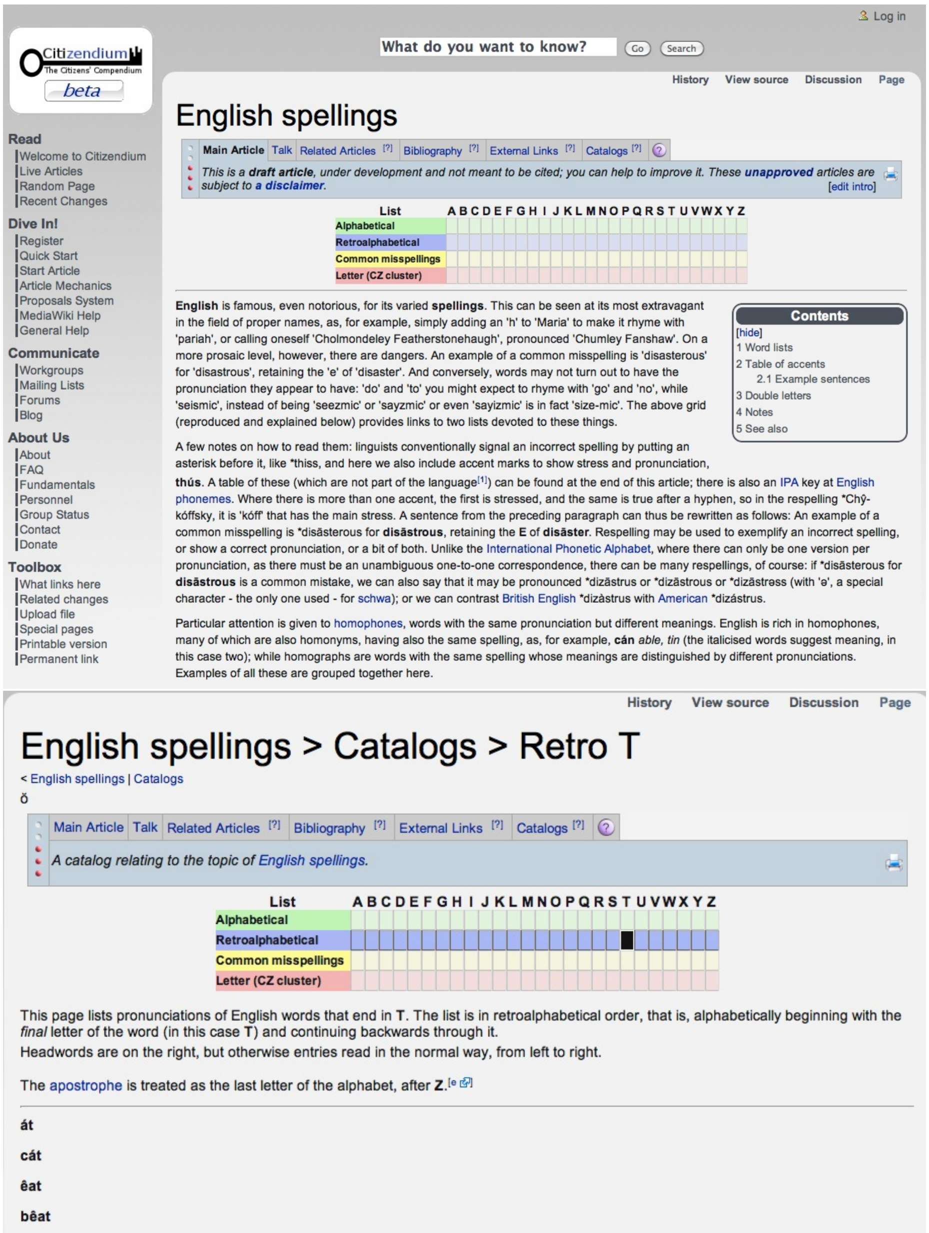
Fig. 2. Top: Screenshot of the Related Articles subpage from the [[Biology]] cluster (which is approved) while logged in. It shows the *Parent topics* and the first section of the *Subtopics* – *subdisciplines*. For each related article, there is a short definition or description of the topic, and a link to its Related Articles subpage (hidden behind the *[r]*), as well as instructions on mouseover and a Table of *Content*. **Bottom:** Related Articles subpage from the [[Open Knowledge Conference]] cluster (while logged out) which has not yet been converted to subpage style but can already be used for structuring information related to the topic. In principle, on could also think of adding [[Open Knowledge Conference 2010]] as a subtopic and using this article for conference blogging. However, the current MediaWiki software cannot handle parallel editing by multiple users, though tools like Etherpad [9] have shown that it is feasible.

the closest analogues from academic papers: All pages are encouraged to have a short Definition subpage (around 30 words or 150 characters) which defines or describes the subject of the page. They are also encouraged to have a comprehensive Related Articles subpage, which uses templates to pull in the definitions from the pages that it links to (a feature that relies on the absence of vandalism). If one looks at the Related Articles subpage of [[Biology]] (cf. Fig. 2, top), one can see the parent topics of biology (science), the subtopics - subdisciplines of biology like zoology, genetics and biochemistry, articles on the history of biology and techniques used by biologists - and finally other related topics, including material on the life cycle, the various biochemical substances like DNA and proteins, the components of the cell, and other specialised language. This Related Articles page gives a pretty comprehensive contextual introduction to what biology is all about, and is structured by the authors of the article in a way that is consistent across the site (cf. Fig. 2, bottom). This goes beyond Wikipedias categories, "See also" sections and ad-hoc infoboxes. Citizendium's approach can be considered as an exploratory next step towards linking encyclopaedic content with the Semantic Web.

Subpages (a further usage example is in (cf. Fig. 3) are one way in which Citizendium is attempting to go beyond what is provided in either traditional paper-based encyclopaedias or by Wikipedia: to engage with context, with related forms of knowledge, and to emancipate knowledge from the page format to which it was confined in the print era. Marx wrote that "Philosophers have hitherto only interpreted the world in various ways; the point is to change it" [11]. Traditional encyclopaedias attempt to reflect the world, but we are attempting to go further. The open science movement - which has formed around the combination of providing open access to journal articles, making scientific data more openly available in raw forms, using and sharing open source software and experimenting with some of the new techniques appearing from the community that is formed under the 'Web 2.0' banner - is exploring the edge of what is now possible for scientists to do to create new knowledge. Some of the electronic engagements by academics has been for actual research benefit, some has just been PR for universities - doing podcasts to sound 'relevant'. The Citizendium model, while a little bit more traditional than some of the open science platforms, is willing to try a variety of new things. Wikipedia has produced a pretty good first version of a collaboratively written encyclopedia – the challenge is to see if we can go further and produce a citizens' compendium of structured and comprehensive knowledge and update it as new evidence or insights arise.

2.4 Open governance

Citizendium has an evolving, but hopefully soon-to-be clearly defined governance process - a Charter is in the process of being drafted by an elected group of writers that will allow for democratic governance and oversight. The broad outline is this: we will have a democratically elected Editorial Council which will deal with content policy and resolving disputes regarding content, and we will also



Read
- Welcome to Citizendium
- Live Articles
- Random Page
- Recent Changes

Dive In!
- Register
- Quick Start
- Start Article
- Article Mechanics
- Proposals System
- MediaWiki Help
- General Help

Communicate
- Workgroups
- Mailing Lists
- Forums
- Blog

About Us
- About
- FAQ
- Fundamentals
- Personnel
- Group Status
- Contact
- Donate

Toolbox
- What links here
- Related changes
- Upload file
- Special pages
- Printable version
- Permanent link

English spellings

Main Article | Talk | Related Articles [?] | Bibliography [?] | External Links [?] | Catalogs [?] | ?

*This is a **draft article**, under development and not meant to be cited; you can help to improve it. These **unapproved** articles are subject to **a disclaimer**.* [edit intro]

List	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Alphabetical																										
Retroalphabetical																										
Common misspellings																										
Letter (CZ cluster)																										

Contents
[hide]
1 Word lists
2 Table of accents
2.1 Example sentences
3 Double letters
4 Notes
5 See also

English is famous, even notorious, for its varied **spellings**. This can be seen at its most extravagant in the field of proper names, as, for example, simply adding an 'h' to 'Maria' to make it rhyme with 'pariah', or calling oneself 'Cholmondeley Featherstonehaugh', pronounced 'Chumley Fanshaw'. On a more prosaic level, however, there are dangers. An example of a common misspelling is 'disasterous' for 'disastrous', retaining the 'e' of 'disaster'. And conversely, words may not turn out to have the pronunciation they appear to have: 'do' and 'to' you might expect to rhyme with 'go' and 'no', while 'seismic', instead of being 'seezmic' or 'sayzmic' or even 'sayizmic' is in fact 'size-mic'. The above grid (reproduced and explained below) provides links to two lists devoted to these things.

A few notes on how to read them: linguists conventionally signal an incorrect spelling by putting an asterisk before it, like *thiss, and here we also include accent marks to show stress and pronunciation, **thús**. A table of these (which are not part of the language[1]) can be found at the end of this article; there is also an IPA key at English phonemes. Where there is more than one accent, the first is stressed, and the same is true after a hyphen, so in the respelling *Chŷ-kóffsky, it is 'kóff' that has the main stress. A sentence from the preceding paragraph can thus be rewritten as follows: An example of a common misspelling is *disāsterous for **disāstrous**, retaining the **E** of **disāster**. Respelling may be used to exemplify an incorrect spelling, or show a correct pronunciation, or a bit of both. Unlike the International Phonetic Alphabet, where there can only be one version per pronunciation, as there must be an unambiguous one-to-one correspondence, there can be many respellings, of course: if *disāsterous for **disāstrous** is a common mistake, we can also say that it may be pronounced *dizāstrus or *dizāstrous or *dizāstrəss (with 'ə', a special character - the only one used - for schwa); or we can contrast British English *dizàstrus with American *dizástrus.

Particular attention is given to homophones, words with the same pronunciation but different meanings. English is rich in homophones, many of which are also homonyms, having also the same spelling, as, for example, **cán** *able, tin* (the italicised words suggest meaning, in this case two); while homographs are words with the same spelling whose meanings are distinguished by different pronunciations. Examples of all these are grouped together here.

English spellings > Catalogs > Retro T

< English spellings | Catalogs

ŏ

Main Article | Talk | Related Articles [?] | Bibliography [?] | External Links [?] | Catalogs [?] | ?

A catalog relating to the topic of English spellings.

List	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Alphabetical																										
Retroalphabetical																				■						
Common misspellings																										
Letter (CZ cluster)																										

This page lists pronunciations of English words that end in **T**. The list is in retroalphabetical order, that is, alphabetically beginning with the *final* letter of the word (in this case **T**) and continuing backwards through it.
Headwords are on the right, but otherwise entries read in the normal way, from left to right.

The apostrophe is treated as the last letter of the alphabet, after **Z**.[e]

át

cát

êat

bêat

Fig. 3. Top: Screenshot of the main page of the [[English spellings]] cluster while logged out. It shows the *blue cluster bar* that indicates that the page has not been approved and links to all the supages. Also visible is the *status indicator* (red dots on the left topped by grey dots, indicating the status of a 'developing' article), and a stronger *disclaimer* than on approved pages. Below this standard header, a set of templates links to 'Catalog' subpages that collect links, for each letter of the English alphabet, to *Alphabetical* and *Retroalphabetical* lists of spellings, to lists of *Common misspellings* as well as to an article on the specific *Letter*. **Bottom:** Close-up of the Catalogs subpage hosting the retroalphabetical list of English spellings for the letter T, again cross-linked with all the other subpages in that cluster.

have a Management Committee, responsible for anything not related to content. The Management Committee appoint Constables who uphold community policy regarding behaviour. Disputes with the Constables can be brought to an Ombudsman selected by the Editorial Council and Management Committee. At the time of writing, the charter is still to be ratified by the community. One of the reasons we have this is that although there is a cost to having bureaucracy and democracy, the benefits of having an open governance process outweigh the costs. We have a real problem when governments of real-life communities are controlled by shadowy cabals who invoke byzantine legal codes - all the same problems would seem to apply to online communities. With a Wikipedia article, the debate seems to shift very quickly from the truth value or relevance of the content itself into often ritualized arguments about acronyms (AfDs, NPOV, CSD, ArbCom, OR etc.). There is always a challenge in any knowledge-based community in attempting to reconcile a fair and democratic process with a meritocratic respect for expertise. There are no easy answers - if we go too far towards bureaucracy, we risk creating a system where management is separated from the actual day-to-day writing of the site, while if we attempt to let the site 'manage itself', we risk creating a rather conservative mob rule that doesn't afford due process to interested outsiders. A more traditional management structure, combined with real names and civility, should help those outside of the online community - the many experts in real life who work in universities, in business and in public life - to participate on an equal footing. Hopefully, if we get the governance decisions right, we can also not get in the way of the people who engage on hobbyist terms with Citizendium.

2.5 Open education

An important part of the governance process is collaboration with partners external to the Citizendium. One of our initiatives – called Eduzendium – provides for educators in higher education to assign work on wiki articles as part of a course. We have most recently had politics students from the Illinois State University work on articles on pressure groups in American public life, as well as medical students from Edinburgh, biologists from City University of New York and the University of Colorado at Boulder, finance students from Temple University and others. These courses reserve a batch of articles for the duration of the course, and assign each article to one or more students. The course instructor can reserve the articles for just the group of students enrolled in the course, or invite the wider Citizendium community to participate. Much of the formatting is achieved via course-specific templates that can be generated semi-automatically by the instructor and applied throughout the course pages, so that course participants can concentrate on content.

3 Open questions

The project is still young, resulting in a number of challenges and opportunities. In many fields, Citizendium does not meet our own standards – we do not have a full range of expert editors. Larry Sanger once envisioned that Citizendium could reach 100,000 articles by 2012. This would, on average, require about 150 new articles a day to reach; the current level is around 15. It is not obvious how the necessary shift from linear to exponential growth can be achieved.

Motivating both editors and authors to take part in both writing and approving of content remains a difficult challenge – most experts have very little time to offer for projects that do not contribute to the metrics according to which their performance is evaluated, and others shy away from contributing under their real name and in the presence of experts. Another problem is that the initial structure of the community, and the nature of its interaction with Wikipedia, has led to a few articles on popular pseudoscientific topics which are hard to handle from an editorial perspective because those willing to invest their time on the topics are usually heavily biased in their approach, and most of those capable of evidence-based comment prefer not to contribute to these topics.

The project also needs to allow for more feedback by non-registered readers, without harming the currently very collegial atmosphere that is to a large extent due to the real-name policy and the respect for expertise. We may need to explore how to codify our core policies and collaboration model as a possible MediaWiki extension, from which other wikis could possibly benefit – online, "code is law" [10], as is currently being highlighted by sites like Stack Overflow which have changed the social interactions of participants by changing formal features of user experience and social structure. We need to find financial backing and support. So far, the project has been run on a basically volunteer-only basis, yet the envisioned growth and improvement of English-language content and the possible start of branches in other languages require a higher degree of professionalisation, for which the upcoming Charter is meant as a basis.

4 Open perspectives

Citizendium is open for partnerships with other open science and online knowledge communities and projects. Possible candidate projects would include, for instance, AcaWiki [1] for references, OpenWetWare for primary research, and Open Access journals [5] as possible content providers, and of course the Wikipedias and other public wikis for exchange on matters of content management, community development and user experience. The key strength we think the Citizendium model brings is a greater focus on knowledge contextualization: it will be interesting to see whether we can evolve the social model for knowledge production to keep up with changes in the technological possibilities. Many in the Citizendium community are looking forward to working alongside both academics and those working in the Semantic Web community to tie Citizendium into data projects. We feel that despite the commoditization of Web 2.0 technologies, there is still plenty of opportunities for reinventing and experimenting with new ways to render and collaborate on knowledge production and to see if we can build a more stable, sustainable and collegial atmosphere – with democratic and meritocratic elements – for experts and the public to work together.

Acknowledgments. The authors wish to thank Russell D. Jones, Howard C. Berkowitz, Steven Mansour and Peter Schmitt for critical comments on earlier versions of this draft as well as Claudia Koltzenburg, François Dongier and Charles van den Heuvel for helpful discussions.

References

1. AcaWiki,
`http://acawiki.org/`
All URLs referenced in this article were functional as of March 31, 2010.
2. Citizendium,
`http://www.citizendium.org/`
3. Clark, D.: Plenary lecture, "A Cloudy Crystal Ball – Visions of the Future", Proc. 24th IETF: 539 (1992),
`http://www.ietf.org/proceedings/prior29/IETF24.pdf`
4. Creative Commons-Attribution-Share Alike license 3.0,
`http://creativecommons.org/licenses/by-sa/3.0/`
5. Directory of Open Access Journals,
`http://www.doaj.org/`
6. Dispersive PDE Wiki,
`http://tosio.math.utoronto.ca/wiki/`
7. Encyclopedia of Cosmos,
`http://www.cosmosportal.org/`
8. Encyclopedia of Earth,
`http://www.eoearth.org`
9. Etherpad source code,
`http://code.google.com/p/etherpad/`
10. Lessig, Lawrence: Code and Other Laws of Cyberspace,
`http://codev2.cc/`
11. Marx, Karl: Theses on Feuerbach,
`http://www.marxists.org/archive/marx/works/1845/theses/theses.htm`
12. OpenWetWare,
`http://www.openwetware.org/`
13. Polymath WIki,
`http://michaelnielsen.org/polymath1/`
14. Priedhorsky R, Chen J, Lam STK, Panciera K, Terveen L, et al. (2007) Creating, destroying, and restoring value in wikipedia. In: GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work. New York, NY, USA: ACM, pp. 259–268. `http://doi.acm.org/10.1145/1316624.1316663`.
15. Raymond, Eric S.: The Cathedral and the Bazaar,
`http://www.catb.org/~esr/writings/homesteading/`
16. RNA Biology, Guidelines for the RNA Families Track,
`http://www.landesbioscience.com/journals/rnabiology/guidelines/`
17. Scholarpedia,
`http://www.scholarpedia.org/`
18. WikiGenes,
`http://www.wikigenes.org/`
19. English Wikipedia,
`http://en.wikipedia.org`

20. Wikiversity,
`http://www.wikiversity.org`
21. Wilbanks, J.: Publishing science on the web,
`http://scienceblogs.com/commonknowledge/2009/07/publishing_science_on_the_web.php`

Utilizing, creating and publishing Linked Open Data with the Thesaurus Management Tool PoolParty

Thomas Schandl, Andreas Blumauer

punkt. NetServices GmbH,
Lerchenfelder Gürtel 43, 1160 Vienna, Austria
schandl@punkt.at, blumauer@punkt.at

Abstract. We introduce the Thesaurus Management Tool (TMT) PoolParty based on Semantic Web standards that reduces the effort to create and maintain thesauri by utilizing Linked Open Data (LOD), text-analysis and easy-to-use GUIs. PoolParty's aim is to lower the access barriers to managing thesauri, so domain experts can contribute to thesaurus creation without needing knowledge about the Semantic Web. A central feature of PoolParty is the enriching of a thesaurus with relevant information from LOD sources. It is also possible to import and update thesauri from remote LOD sources. Going a step further we present a Personal Information Management tool built on top of PoolParty which relies on Open Data to assist the user in the creation of thesauri by suggesting categories and individuals retrieved from LOD sources. Additionally PoolParty has natural language processing capabilities enabling it to analyse documents in order to glean new concepts for a thesaurus and several GUIs for managing thesauri varying in their complexity. Thesauri created with PoolParty can be published as Open Knowledge according to LOD best practices.

Keywords: Semantic Web, Linking Open Data, Thesaurus, Personal Information Management, SKOS, RDF.

1 Introduction

Thesauri have been an important tool in Information Retrieval for decades and still are [1]. While they have the potential to greatly improve the information management of organisations, professionally managed thesauri are rarely used in content management systems, search engines or tagging systems.

Important reasons frequently given for this are: (1) the difficulty of learning and using TMT, (2) the lacking possibilities to integrate TMTs into existing information systems, (3) it is laborious to create and maintain a thesaurus, and while TMTs often support either automatic or manual methods to maintain a thesaurus they rarely combine those two approaches, and (4) users don't have enough knowledge about thesaurus building methodologies and/or worthwhile use cases utilizing semantic knowledge models like SKOS thesauri.

The TMT PoolParty[1] addresses the first three issues. A central goal is to ease the process of creating and maintaining thesauri by domain experts, that don't have a strong technical background, don't know about semantic technologies and maybe know little about thesauri. We see an important role for Linked Open Data in this area and equipped PoolParty with the capability to enrich one's own knowledge model with relevant information from the LOD cloud. In combination with several GUIs suited for varying levels of complexity PoolParty allows for low access barriers for creating and utilizing thesauri and Open Data.

PoolParty is a commercial application, but will have a version that can be used free of charge. We are working on such a version that makes use of the Talis platform[2]. In any version the user will have the option to publish thesauri as LOD and license them under various Creative Commons licenses.

2 Use Cases

PoolParty is based on Semantic Web technologies like RDF[3] and SKOS[4] (Simple Knowledge Organisation System) allowing for multilingual thesauri to be represented in a standardised manner [2]. While OWL[5] would offer greater possibilities in creating knowledge models, it is deemed too complex for the average information worker.

PoolParty was conceived to facilitate various commercial and non-commercial applications for thesauri. In order to achieve this, it needs to publish them and offer methods of integrating them with various applications [3]. In PoolParty this can be realized on top of its RESTful web service interface providing thesaurus management, indexing, search, tagging and linguistic analysis services.

Some of these (semantic) web applications are:

- Semantic search engines
- Recommender systems (similarity search)
- Corporate bookmarking
- Annotation- & tag recommender systems
- Autocomplete services and facetted browsing.
- Personal Information Management

These use cases can be either achieved by using PoolParty stand-alone or by integrating it with existing (Enterprise) Search Engines and Document Management Systems.

[1] http://poolparty.punkt.at/
[2] http://www.talis.com/platform/
[3] http://www.w3.org/RDF/
[4] http://www.w3.org/2004/02/skos
[5] http://www.w3.org/TR/owl-ref/

3 Technologies

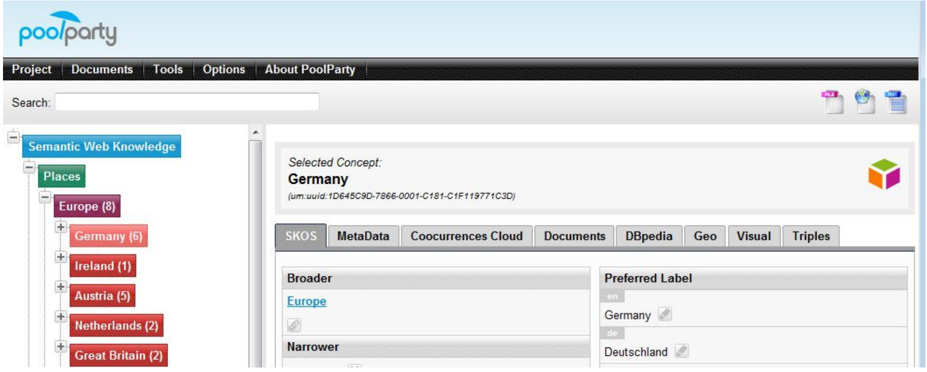
PoolParty is written in Java and uses the SAIL API[6], whereby it can be utilized with various triple stores, which allows for flexibility in terms of performance and scalability.

Thesaurus management itself (viewing, creating and editing SKOS concepts and their relationships) can be done in an AJAX Frontend based on Yahoo User Interface (YUI). Editing of labels can alternatively be done in a Wiki style HTML frontend.

For key-phrase extraction from documents PoolParty uses a modified version of the KEA[7] 5 API, which is extended for the use of controlled vocabularies stored in a SAIL Repository (this module is available under GNU GPL). The analysed documents are locally stored and indexed in Lucene[8] along with extracted concepts and related concepts.

4 Thesaurus Management with PoolParty

The main thesaurus management GUI of PoolParty (see Fig. 1) is entirely web-based and utilizes AJAX to e.g. enable the quick merging of two concepts either via drag & drop or autocompletion of concept labels by the user. An overview over the thesaurus can be gained with a tree or a graph view of the concepts.



Selected Concept:
Germany
(urn:uuid:1D645C9D-7866-0001-C181-C1F119771C3D)

Broader
Europe

Narrower

Preferred Label
en
Germany
de
Deutschland

Fig. 1 PoolParty's main GUI with concept tree and SKOS view of selected concept

Consistent with PoolParty's goal of relieving the user of burdensome tasks while managing thesauri doesn't end with a comfortable user interface: PoolParty helps to semi-automatically expand a thesaurus as the user can use it to analyse documents (e.g. web pages or PDF files) relevant to her domain in order to glean candidate terms for her thesaurus. This is done by the key-phrase extractor of KEA. The extracted

[6] http://www.openrdf.org/doc/sesame2/system/ch05.html
[7] http://www.nzdl.org/Kea/index.html
[8] http://lucene.apache.org/

terms can be approved by the user, thereby becoming "free concepts" which later can be integrated into the thesaurus, turning them into "approved concepts".

Documents can be searched in various ways – either by keyword search in the full text, by searching for their tags or by semantic search. The latter takes not only a concept's preferred label into account, but also its synonyms and the labels of its related concepts are considered in the search. The user might manually remove query terms used in semantic search. Boost values for the various relations considered in semantic search may also be adjusted. In the same way the recommendation mechanism for document similarity calculation works.

PoolParty by default also publishes an HTML Wiki version of its thesauri, which provides an alternative way to browse and edit concepts. Through this feature anyone can get read access to a thesaurus, and optionally also edit, add or delete labels of concepts. Search and autocomplete functions are available here as well.

The Wiki's HTML source is also enriched with RDFa, thereby exposing all RDF metadata associated with a concept to be picked up the RDF search engines and crawlers.

PoolParty supports the import of thesauri in SKOS (in serializations including RDF/XML, N-Triples or Turtle) or Zthes format.

6 Linked Open Data Capabilities

PoolParty not only publishes its thesauri as Linked Open Data (additionally to a SPARQL endpoint), but it also consumes LOD in order to expand thesauri with information from LOD sources. Concepts in the thesaurus can be linked to e.g. DBpedia[9] via the DBpedia lookup service [4], which takes the label of a concept and returns possible matching candidates. The user can select the DBpedia resource that matches the concept from his thesaurus, thereby creating an owl:sameAs relation between the concept URI in PoolParty and the DBpedia URI. The same approach can be used to link to other SKOS thesauri available as Linked Data.

Other triples can also the retrieved from the target data source, e.g. the DBpedia abstract can become a skos:definition and geographical coordinates can be imported and be used to display the location of a concept on the map, where appropriate. The DBpedia category information may also be used to retrieve additional concepts of that category as siblings of the concept in focus, in order to populate the thesaurus.

PoolParty is not only capable of importing a SKOS thesaurus from a Linked Data server, it may also receive updates to thesauri imported this way. This feature has been implemented in the course of the KiWi[10] project funded by the European Commission. KiWi also contains SKOS thesauri and exposes them as LOD. Both systems can read a thesaurus via the other's LOD interfaces and may write it to their own store. This is facilitated by special Linked Data URIs that return e.g. all the top-concepts of a thesaurus, with pointers to the URIs of their narrower concepts, which allow other systems to retrieve a complete thesaurus through iterative dereferencing of concept URIs.

[9] http://dbpedia.org/

[10] http://kiwi-project.eu/

Additionally KiWi and PoolParty publish lists of concepts created, modified, merged or deleted within user specified time-frames. With this information the systems can learn about updates to one of their thesauri in an external system. They then can compare the versions of concepts in both stores and may write according updates to their own store.

This means each system decides autonomously which data it accepts and there is no risk of a system pushing data that might lead to inconsistencies into an external store. Data transfer and communication are achieved using REST/HTTP, no other protocols or middleware are necessary. Also no rights management for each external systems is needed, which otherwise would have to be configured separately for each source.

The synchronisation process via Linked Data will be improved in the ongoing KiWi project. We will implement an update and conflict resolution dialogue through which a user may decide which updates to concepts to accept and to consequently write to the system's store.

7 Personal Information Management utilizing Linked Open Data

An example application that we are currently developing on top of PoolParty web services is a Personal Information Manager (PIM) utilizing Open Data.

Our goal is to enable users to create and utilize thesauri without requiring any knowledge about thesauri. We aim to hide the complexity of thesauri and their poly-hierarchical structure and concentrate on presenting the user with listboxes filled with terms from a thesaurus or LOD sources.

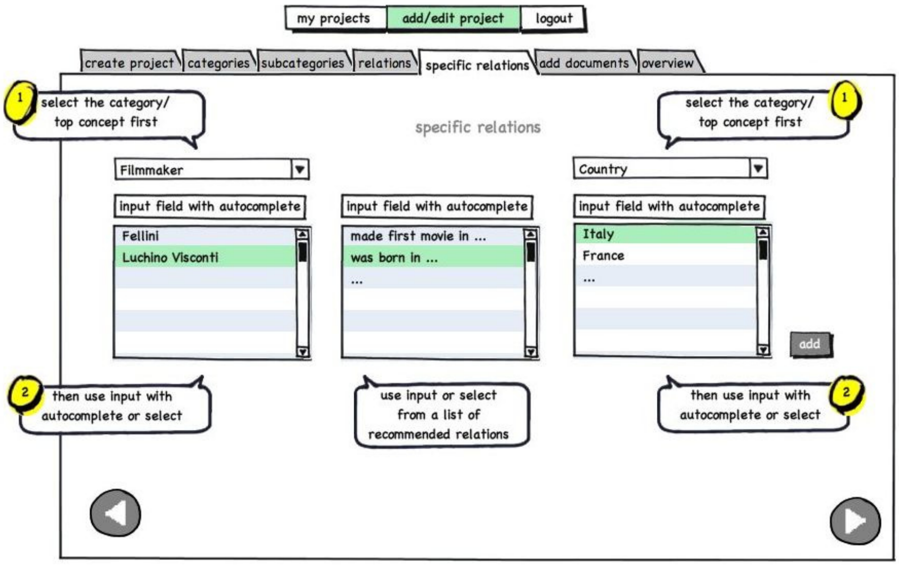
The PIM will be a web based application that makes use of Linked Open Data in order to assist the users with suggestions when they e.g. create categories. A movie expert for example might want to create a knowledge model of filmmakers and the countries they lived in. Upon the creation of a new project, the PIM asks the user to specify a general domain it is about (people, places, things, events, organisations, etc.). After the user selects "people", the system can use e.g. data about categories from YAGO[11], UMBEL[12] or DBpedia that relate to "people" to help refine the user's domain. The system might suggest popular categories or the user can start entering a string like "fil" prompting the system's autocomplete functionality to suggest the DBpedia class "filmmaker". When the user confirms that this is one of the topics his project is about and finishes the same process for the other topics (i.e. countries), several links between the local model and the LOD cloud exist and more specific information can be retrieved to assist the user's work on the thesaurus. The system might suggest possible relations between painters and cities like "made movie in", "was born in" or "lived in", and the user can specify which particular relations are of interest to him.

In a similar way the PIM will assist with creating instance data and suggest particular filmmakers and countries (see mock-up in Fig. 2) from sources like DBpedia that belong to the corresponding classes. In this way the PIM not only helps

[11] http://www.mpi-inf.mpg.de/yago-naga/yago/

[12] http://www.umbel.org/

with rapidly filling the model, but it automatically interlinks it with the LOD cloud in one go.


my projects
add/edit project
logout
create project
categories
subcategories
relations
specific relations
add documents
overview
1 select the category/ top concept first
specific relations
select the category/ top concept first 1
Filmmaker
Country
input field with autocomplete
input field with autocomplete
input field with autocomplete
Fellini
Luchino Visconti
made first movie in ...
was born in ...
...
Italy
France
...
add
2 then use input with autocomplete or select
use input or select from a list of recommended relations
then use input with autocomplete or select 2

The user will also be able add his own classes and instances or use PoolParty's natural language processing service to analyse web pages or documents to glean new concepts for use in the model.

Of course this PIM will not only consume LOD, but it can also publish the user created knowledge models as part of the LOD cloud. In this way LOD can be harnessed to enable the average internet user to create more Open Knowledge. There will be an online version of this PIM that can be used free of charge.

In the upcoming project LASSO funded by the Austrian Research Promotion Agency (FFG)[13] we will do research on algorithms that enable smart interlinking of local data and LOD sources, which will be used for the PIM. Amongst the algorithmic solutions we will pursue are graph based look-up services (e.g. querying LD sources by taking context into account instead of just searching for keywords), network search methods like Spreading Activation and statistical methods such as the Principal Component Analysis.

8 Final Remarks

We have shown how Open Linked Data can help in various ways with easing the creation of knowledge models like thesauri. At OKCon 2010 we will demonstrated PoolParty and session visitors will learn how to manage a SKOS thesaurus and how

[13] http://ffg.at/content.php

PoolParty supports the user in this process. The document analysis features will be presented, showing how new concepts can be gleaned from text and integrated into a thesaurus.

It will be shown how to interlink local concepts from DBpedia, thereby enhancing one's thesaurus with triples from the LOD cloud. Finally the state of the PoolParty PIM tool will be presented.

References

1. Aitchison, J., Gilchrist, A., Bawden, D.: Thesaurus Construction and Use: A Practical Manual. 4th edn. Europa Publications (2000)
2. Pastor-Sanchez, J. P., Martínez Mendez, F., and Rodríguez-Muñoz, J. V.: Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. informationresarch Vol 14 No. 4, Dec 2009. http://informationr.net/ir/14-4/paper422.html
3. Viljanen, K., Tuominen, J., Hyvönen, E.: Publishing and using ontologies as mashup services. In: Proceedings of the 4th Workshop on Scripting for the Semantic Web (SFSW 2008), 5th European Semantic Web Conference 2008 (ESWC 2008), Tenerife, Spain (June 1-5 2008)
4. Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S.: DBpedia - A crystallization point for the Web of Data. Web Semantics: Science, Services and Agents on the World Wide Web. Volume 7, Issue 3, September 2009, Pages 154-165